

---

# BACHELOR THESIS

---

Victoria Pratzka

***In silico* Analysis of Genes Involved  
in the Initiation of Barley Pollen  
Embryogenesis**

Mittweida, 2013

## **BACHELOR THESIS**

---

# ***In silico* Analysis of Genes Involved in the Initiation of Barley Pollen Embryogenesis**

author:

**Victoria Pratzka**

course of studies:

**Biotechnology/Bioinformatics**

seminar group:

**BI10w2-B**

first examiner:

**Prof. Dr. rer. nat. Röbbbe Wünschiers**

second examiner:

**Dr. Uwe Scholz**

submission:

**Mittweida, 21.08.2013**

defence:

**Mittweida, 26.08.2013**

**Bibliographic Description:**

Pratzka, Victoria: *In silico* Analysis of Genes Involved in the Initiation of Barley Pollen Embryogenesis. - 2013 - 13, 54, 21 S. Mittweida, Hochschule Mittweida - University of Applied Science, Faculty MNI, Bachelor Thesis, 2013

**Abstract:**

The main purpose of this Bachelor thesis was to find and to compile comprehensive information on barley genes expressed in the context of pollen embryogenesis. In the present study, this approach was confined to genes that were previously known to be associated with the initiation of embryogenesis in different plant species. First, candidate transcript sequences were identified in barley. Second, transcript and associated genomic sequences were analyzed *in silico* to provide suitable structural and functional annotations. Finally, the results of one representative example are presented and interpreted in detail. This work aims to contribute to a significantly improved understanding of pollen embryogenesis - a biological phenomenon broadly used for haploid technology in crop improvement.

## **Acknowledgements**

The present Bachelor thesis has been carried out at the Leibniz Institute of Plant Genetics and Crop Plant Research (IPK) in Gatersleben in the research groups of Bioinformatics and Information Technology (BIT) and Plant Reproductive Biology (PRB).

First I would like to thank my supervisors Dr. Uwe Scholz and Dr. Jochen Kumlehn for giving me the opportunity to do this Bachelor thesis in their groups. Both inspired me to bring the best out of this Bachelor thesis.

I thank Dr. Uwe Scholz for the many fruitful discussions, his continuous support for my work and his efforts in correcting and optimizing this thesis.

I am also very grateful to Dr. Jochen Kumlehn for his patience and his valuable discussions which helped me to find the right direction for this work.

Furthermore my thanks go to the laboratory assistance Mrs. Ingrid Otto, from the PRB group, for giving me the opportunity to try myself on extracting stress-induced pollen from anthers and observing them run the different stages of pollen embryogenesis.

In addition I thank M.Sc. Sebastian Beier for his patient and helpful briefing in the command line implementations and all further informatical issues. Moreover I am thankful about his efforts in correcting my internship report and this present thesis.

Second to last I thank Prof. Dr. rer. nat. Röbbke Wünschiers for his attendance to support me as supervisor from the HS Mittweida.

Finally I thank Tino Kreszies for his backup, encouragement and bearing of any mood all the time.

**Statement of Authorship**

I hereby certify that this Bachelor thesis presented here has been composed by myself and is the result of my own investigations, unless otherwise acknowledged in the text. All references and all sources of information have been specifically acknowledged. This thesis has not been submitted, either in part or whole, for a degree at this or any other University. This work has not been published.

Mittweida, 21<sup>th</sup> August 2013

Victoria Pratzka

**Index of Contents**

<b>Acknowledgements.....</b>	<b>I</b>
<b>Statement of Authorship.....</b>	<b>II</b>
<b>Index of Contents.....</b>	<b>III</b>
<b>List of Figures.....</b>	<b>V</b>
<b>List of Tables.....</b>	<b>VI</b>
<b>List of Abbreviations .....</b>	<b>VII</b>
<b>1 Introduction .....</b>	<b>1</b>
1.1 Motivation .....	1
1.2 Purpose .....	1
1.3 Outline .....	3
<b>2 Fundamentals .....</b>	<b>4</b>
2.1 Barley as a Crop and a Model Organism.....	4
2.2 The Whole Genome Shotgun Assembly of Barley .....	5
2.3 Totipotency of Plant Cells and Haploid Technology .....	6
2.4 Pollen Embryogenesis.....	7
2.5 RNA-Seq Technology.....	9
2.5.1 Basic Protocol of RNA-Seq .....	10
2.5.2 Benefits and Challenges of RNA-Seq .....	10
2.5.3 The RNA-Seq Data Set.....	12
<b>3 Methods .....</b>	<b>14</b>
3.1 Processing of the Data Set .....	14
3.2 Analysis Steps.....	14
3.3 BLAST.....	18
3.3.1 BLASTn .....	19
3.3.2 BLASTx.....	20
3.4 ClustalW2 .....	20
3.5 Conserved Domain Database.....	22
3.6 Mulan .....	23

3.7 TriAnnot Pipeline .....	24
3.8 Tablet .....	25
3.9 Statistical Evaluation.....	29
<b>4 Results .....</b>	<b>32</b>
4.1 Structural Gene Annotation.....	33
4.2 Functional Gene Annotation.....	39
4.3 Summarization .....	45
<b>5 Discussion .....</b>	<b>49</b>
<b>6 Summary and Outlook.....</b>	<b>52</b>
6.1 Summary .....	52
6.2 Outlook .....	53
<b>List of References.....</b>	<b>VIII</b>
<b>Appendix .....</b>	<b>XIV</b>
Appendix 1 .....	XIV
Appendix 2 .....	XXVII
Appendix 3 .....	XXVIII
Appendix 4 .....	XXIX

## **List of Figures**

Figure 1 - Collage of Barley [URL-12/-13/-14/-15].....	4
Figure 2 - Pollen Development [edited after Maraschin, S. F. et al. (2006)] .....	8
Figure 3 - RNA-Seq Basic Protocol [edited after Wang, Z. et al. (2009); URL-10/-11]....	11
Figure 4 - Activity Diagram of the Analysis Steps .....	17
Figure 5 - GBrowse Viewer of TriAnnot .....	25
Figure 6 - Interface of Tablet .....	27
Figure 7 - Main Display of Tablet Showing Properly Paired Reads.....	28
Figure 8 - Main Display of Tablet Showing a Read with Unmapped Mate.....	28
Figure 9 - Read Count Normalization [edited after Garber, M. et al. (2011)] .....	29
Figure 10 - Multiple-Sequence Local Alignment of Mulan .....	34
Figure 11 - Transcription Factor Binding Sites Profile of MultiTF .....	35
Figure 12 - Morex WGS-Contig Visualization in the GBrowse Viewer of TriAnnot .....	37
Figure 13 - Read Coverage Graph of the Morex WGS-Contig in Tablet .....	39
Figure 14 - Conserved Domain Database View .....	40
Figure 15 - Conserved Region in the Multiple Protein Sequence Alignment .....	41
Figure 16 - Distance Tree according to the Protein MSA .....	43
Figure 17 - Startcodon Detection with Tablet .....	44
Figure 18 - Read Coverage of the RNA-Seq-Contig.....	45



**List of Tables**

Table 1 - Summary of the Methods .....	30
Table 2 - Summarization of the Achieved Results .....	46
Table 3 - Conjugated Genes with Same Sequences.....	47

**List of Abbreviations**

ABI3	Abscisic Acid Insensitive 3
ANAP	Arabidopsis Network Analysis Pipeline
BAC	Bacterial Artificial Chromosome
BLAST	Basic Local Alignment Search Tool
BLASTn	BLAST search in nucleotide database using a nucleotide query
BLASTx	BLAST search in protein database using a translated nucleotide query
bp	base pair(s)
CDD	Conserved Domain Database
COG	Clusters of Orthologous Groups
DNA	Deoxyribonucleic Acid
ECR	Evolutionary Conserved Regions
EST	Expressed Sequence Tags
FPKM	Fragments Per Kilobase of exon per Million fragments mapped
GALA	Genome Alignment and Annotation Database
Gb	Giga base pairs
Kb	Kilo base pairs
Mb	Mega base pairs
MSA	Multiple Sequence Alignment
MSLA	Multiple-Sequence Local Alignment
NCBI	National Center for Biotechnology Information
NGS	Next Generation Sequencing
PCR	Polymerase Chain Reaction
Pfam	Protein Families
RNA	Ribonucleic acid
RNA-Seq	RNA Sequencing Technology
RPS-BLAST	Reverse Position Specific BLAST
SMART	Simple Modular Architecture Research Tool
SNP	Single Nucleotide Polymorphism

SRA	Sequence Read Archive
TBA	Threaded Blockset Aligner
TF	Transcription Factor
TFBS	Transcription Factor Binding Sites
WGS	Whole Genome Shotgun

## **1 Introduction**

### **1.1 Motivation**

The phenomenon of pollen embryogenesis is a survival adaption of plants which do not occur in nature regularly. It is an impressive example of totipotency of plants which was published first in 1964. In this publication it is described that haploid plants can be received with culturing anthers under specific *in vitro* conditions. Confirmed by further researches it could be observed that the microspores within the anthers run an alternative development and give rise to completely new haploid, but sterile plants [Reynolds, T. L. (1997); Silva, T. D. (2012)].

This led to the haploid technology which is in main interest of plant breeders, researchers as well as plant scientists. Haploid plants can give rise to perfectly homozygous fertile plants through chromosome duplication. This is especially important for plant breeding because with the haploid technology there is no need for manual selective breeding across numerous generations. And plant scientists expect to find new information, about the basic pathways and interactions of embryogenic development, for both somatic as well as zygotic embryogenesis, from investigations of pollen embryogenesis. But in spite of many investigations only the basic principles of pollen embryogenesis could be explained yet. And there are only a few genes and proteins detected that are involved verifiable in the transition from regular pollen development to pollen embryogenesis.

### **1.2 Purpose**

The *in silico* analyses that have been executed for the present Bachelor thesis should provide appropriate putative homologs in barley (*Hordeum vulgare*) to already known genes associated with the initiation of embryogenesis in different related plant species like thale cress (*Arabidopsis thaliana*), rice (*Oryza sativa*) or wheat (*Triticum aestivum*) on one hand. And on the other hand these investigations should provide a

bioinformatic basis for further molecular biological analyses of barley pollen embryogenesis and contribute to an improved understanding of pollen embryogenesis.

The initial point was the sequencing of the transcriptome of three states of microspores crucial for the initiation of pollen embryogenesis. These three stages represent the switch from regular development to pollen embryogenesis. And because of the huge amount provided by the high-throughput sequencing technique (RNA-Seq) the transcriptomic data sets have to be analyzed *in silico*.

After various general analyses there is a need for some more detailed and therefore manual investigations to execute arising evaluation task settings. A detailed analysis, especially for the present data set, can be performed by two different approaches due to different purposes. In this present Bachelor thesis only one specific approach was followed with the consequent task settings:

- 1) Find appropriate analog genes in barley (*Hordeum vulgare*) to a previous selected list of candidate genes from an educated guess.
- 2) Detect suitable methods and tools to analyze these genes of barley *in silico* and create a pipeline as an instruction manual by which these analyses can be executed.
- 3) Annotate the corresponding genes in barley structurally and functionally with the available bioinformatic methods and tools.

The main aim of this Bachelor thesis was to annotate a selection of candidate genes and their associated putative homologs in barley to achieve information about several features: (1) exonic structure, (2) transcription factor binding sites, (3) start and end positions, (4) conserved regions, (5) repetitive regions, (6) coverage of transcripts and (7) protein sequence, also in comparison with sequences from related plant species.

### 1.3 Outline

At first the present Bachelor thesis gives some fundamentals to provide detailed information about the most important issues. The regarded plant, barley (*Hordeum vulgare*), is introduced as a crop plant and as a model organism which is significant for both nutritional purposes and scientific applications. Furthermore the Morex WGS assembly is presented because it is the major reference of the *in silico* analyses. Afterwards the principles of the totipotency and the phenomenon of pollen embryogenesis and concerning biological basics are roughly explained. Moreover the RNA-Seq technology was depicted. This is the technique by which the transcriptome data of the three stages of microspore development was accumulated. And at last the data sets provided by the RNA-Seq runs were introduced as a reference and as a basis for the statistical evaluation and expression pattern observation.

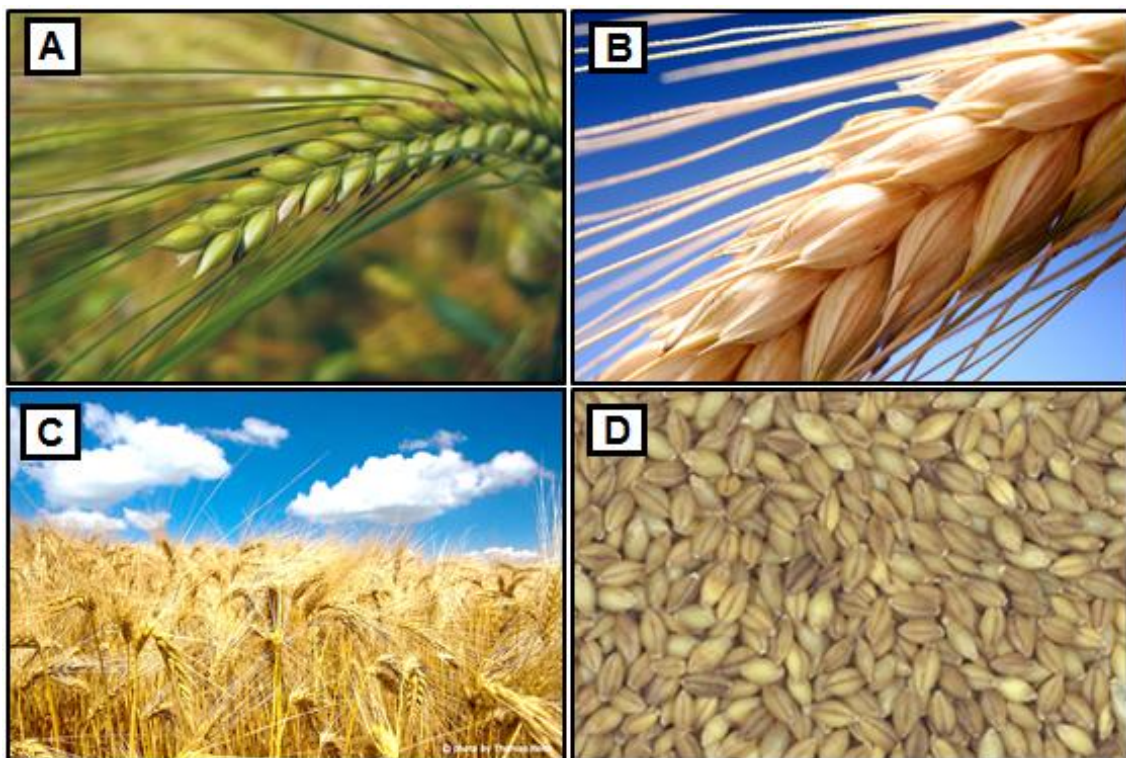
In the second part all used methods and bioinformatic tools are characterized. Therefore the main intention of the tool and the particular implementation for the present task settings is described to get a general idea about the utilization. All these tools were processed in a specific hierarchy that needed to be maintained advisedly. This hierarchy was drafted as an UML activity diagram to give an instruction manual which should help the following readers or users to comprehend the analysis and to apply the methods themselves.

As third part the results achieved while all analyses were presented and interpreted with a suitable representative example out of the previously selected list of candidate genes. The results are divided into a structural and a functional gene annotation. All obtained features are kept digital to give the opportunity to reproduce the results and also have a more specialized view if required. In the end all results of all analyzed genes are condensed in tables to give an overview about the available features.

## **2 Fundamentals**

### **2.1 Barley as a Crop and a Model Organism**

Barley (*Hordeum vulgare* L.) (see Figure 1) is a member of the tribe Triticeae within the grass family Poaceae and belongs to the genus *Hordeum*. It is one of the earliest domesticated crop plants in the world and represents the fourth most abundant cereal after wheat (*Triticum aestivum*), maize (*Zea mays*) and rice (*Oryza sativa*). As a crop species barley emphasizes particular importance because it is widely adapted to variable environmental conditions. In addition barley is much more stress tolerant than wheat and that's why it remains a major food source in poorer countries. The range of use implies mainly the animal feed, the human food and the malt production [The International Barley Genome Sequencing Consortium (2012)].



**Figure 1 - Collage of Barley [URL-12/-13/-14/-15]**

(A) A close-up view of the spike of the common barley (*Hordeum vulgare*) [URL-12]. (B) A detailed view of maturated spikes of barley [URL-13]. (C) A grainfield with mature barley [URL-14]. (D) Grains of barley after harvesting [URL-15].

Beside its importance as a nutritional source barley was established by plant scientists as a model organism. Barley is used for basic genetic research because it is a diploid and temperate plant. Additionally a collection of mutants is available containing most of the morphological and developmental variations of barley. Traditionally barley is considered a model organism for genetic research because it was used for investigations that provide the basis for population and evolutionary genetics. Today a public collection of ESTs (Expressed Sequence Tags), a huge genomic BAC (Bacterial Artificial Chromosome) library and an Affymetrix microarray for crop plants is available [URL-1]. Furthermore a great number of raw sequencing data from various next generation sequencing (NGS) platforms are stored in the Sequence Read Archive (SRA) allocated by the National Center for Biotechnology Information (NCBI) [URL-16]. In addition there is also an amount of sequence-verified single nucleotide polymorphisms (SNPs) which were used for the establishment of a high-throughput SNP genotyping platform based on Illumina [URL-1].

## **2.2 The Whole Genome Shotgun Assembly of Barley**

But while all investigations in barley the major drawback is the absence of an entire reference genome sequence. Therefore the International Barley Genome Sequencing Consortium provided an appropriate sequence reference in form of the barley whole genome shotgun (WGS) assembly in November 2012 [The International Barley Genome Sequencing Consortium (2012)]. This assembly is connected to a genome-wide physical map of the barley cultivar Morex which provides access to the majority of genes from barley and is now an essential reference for genetic research and plant breeding.

With a whole-genome size of 5.1 Gb of barley the assembly represents approximately 95% in the physical map. The WGS assembly consists of 2.670.738 contigs which assemble to a total contig size of 1.868.648.155 bp. Pursuant to this the WGS assembly represents approximately 1.9 Gb of the whole length of the barley genome. The main limiting problem of the physical map is the huge amount of repetitive DNA influencing



the WGS assembling. This repetitive DNA consists of retrotransposons, mobile elements and other repeat structures. Due to this, a noticeable part of the shotgun data collapsed into small contigs which were detected by the outstanding high read depth [The International Barley Genome Sequencing Consortium (2012)]. With the publication of the barley WGS the global barley community makes an important foundation for further and more specific investigations available. In fact this WGS of barley also was the main reference while all *in silico* analysis steps for this bachelor thesis. The parts of the WGS are stored as Morex WGS-Contigs, meaning contiguous sequences that were achieved by the assembling. These Morex WGS-Contigs were used for the *in silico* analysis as genomic reference sequences of barley.

### **2.3 Totipotency of Plant Cells and Haploid Technology**

Plant cells distinguish oneself due to their extraordinary potential for totipotency. This becomes noticeable with the ability of any differentiated plant cell to return to embryogenic development and regenerate a new plant. This competence is supposed to be one of the most important survival adaption of plants. One impressive example of totipotency is the phenomenon termed pollen embryogenesis (also called microspore embryogenesis or androgenesis) [Reynolds, T. L. (1997)].

In pollen embryogenesis the basic principle of totipotency gives a microspore (pollen) the ability to switch from its destined gametogenic development to embryogenic development under specific conditions (see Figure 2) [Silva, T. D. (2012)]. The first report about pollen embryogenesis was published in 1964 from Guha and Maheshwari who cultured anthers of *Datura innoxia* receiving haploid plants. Subsequent to this effort the implementation of cultured anthers for achieving haploid plants has been published for more than 170 species [Reynolds, T. L. (1997)].

Unfortunately haploid plants are weak and sterile. Consequently they did not serve any useful purpose by themselves. But by duplicating their chromosome number diploid plants can be received, which are perfectly homozygous. And these perfectly

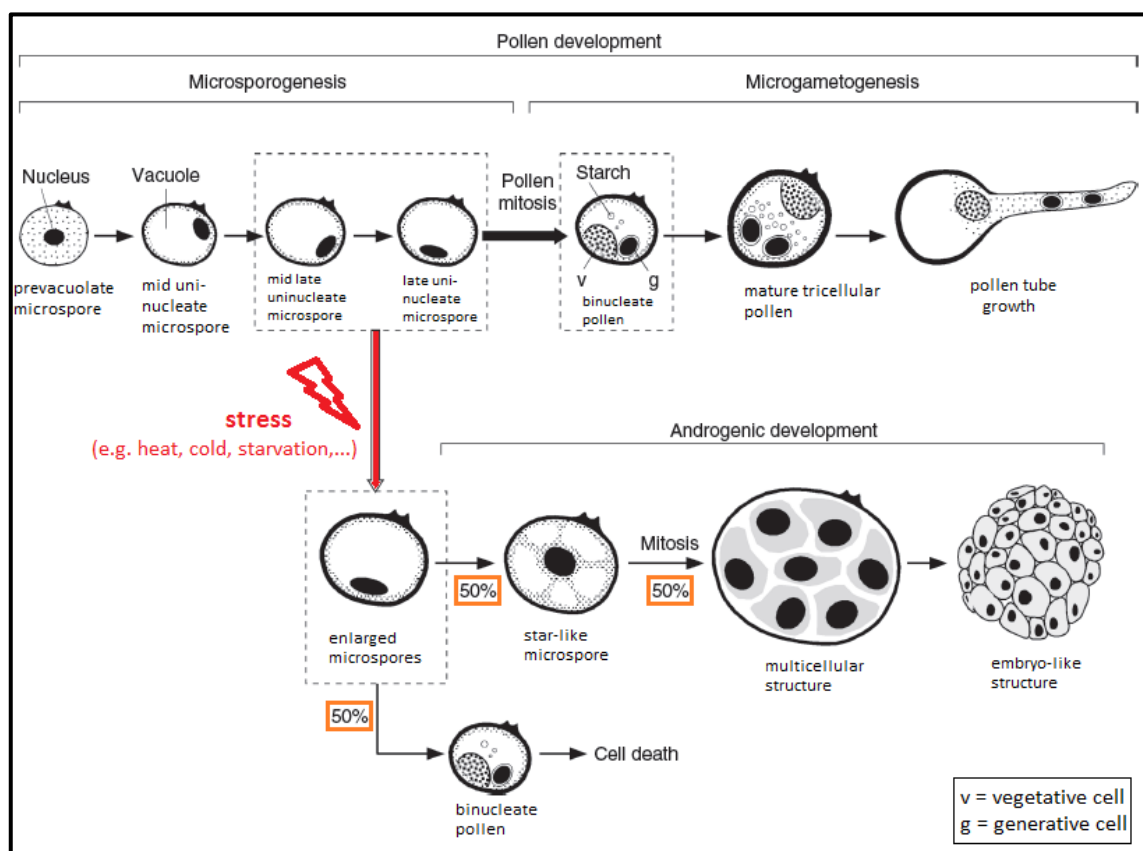
homozygous, diploid plants will give rise to fertile homogenous progenies afterwards (already in the second-generation) [Silva, T. D. (2012)]. Homozygous plants are important for plant breeding as well as for basic research. The technique of pollen embryogenesis gives rise for the haploid technology and serves an enormous capacity for plant scientists. As it can be seen in Figure 2 pollen develops into an embryo-like structure with a given probability. This value differs between the species, pollen embryogenesis was observed for. In barley about 30 % of the induced microspores run pollen embryogenesis in the end, while the rest is dying. At this point the investigations get actually entitled, because the ambition of plant scientists is to increase the percentage and number of microspores that run pollen embryogenesis.

## **2.4 Pollen Embryogenesis**

Microspores are the preliminary stage of the male gametes in plants and regularly develop within the anthers into pollen grains. Regular pollen ontogeny can be divided into two phases (see Figure 2). The first phase is termed microsporogenesis and displays the development of the immature pollen within the anthers by meiosis of the mother cells. An asymmetric mitotic division leads to the microgametogenesis which is the second period of the pollen ontogeny. The mitotic division generates two unequal sized cells. The larger cell is termed vegetative cell and the smaller one is called the generative cell [Lippmann, R. (2012)]. The latter runs another mitotic division which produces two non-flagellated male gametophytes (sperm cells) which are involved in double fertilization later [Reynolds, T. L. (1997)].

In contrast to regular pollen ontogeny, pollen embryogenesis runs in three steps: (1) application of stress, (2) development of multi-cellular structures and (3) formation of an embryogenic structure. In comparison with the regular pollen ontogeny the embryogenic development is initiated by a symmetric mitotic division instead of an asymmetric division. Through further mitotic divisions other dedifferentiated cells arise and because of the rapid proliferation a multi-cellular mass accumulates (see Figure 2). This multi-cellular mass is able to develop into an embryogenic structure, or

it stays a meristem. There are some marker which can be helpful in detecting embryogenic pollen: (a) partitioning of the vacuole, (b) re-positioning of the nucleus, (c) enlargement of the cell volume, (d) generating of a new cell wall, (e) decreasing of the nucleus, (f) degeneration of the plastids, (g) symmetric cell division, (h) star-like structure of the cell, (i) storage reduction of starch and lipids, (j) re-arrangements of the cytoskeleton [Lippmann, R. (2012)]. These markers are able to indicate the androgenic development but there are also some cells which run pollen embryogenesis without showing all the possible markers.



**Figure 2 - Pollen Development [edited after Maraschin, S. F. et al. (2006)]**

Illustration of the stages of pollen while regular pollen- and stress-induced androgenic development in barley. The regular pollen development is divided into microsporogenesis and microgametogenesis. The bold black arrow marks the switch to microgametogenesis. And the bold red arrow marks the transition to an alternative development. The androgenic development starts after the induction of abiotic stress.

As it can be seen in Figure 2, pollen develops into an embryo-like structure with a given probability which differs between the species. According to experiences of the

laboratory assistances from the IPK, about 30 % of the induced microspores in barley run pollen embryogenesis in the end, while the rest is dying.

Pollen embryogenesis rarely occurs in nature but is an adaptive mechanism for survival which can effortlessly be induced under specific *in vitro* conditions [Silva, T. D. (2012)]. For many species those conditions were detected so far. But every species (and also cultivar) seems to be sensitive for species-specific conditions.

In general, the transition from microsporogenesis to microgametogenesis in regular pollen development is a very sensitive period and seems to be the most important window for an induced switch to embryogenic development [Reynolds, T. L. (1997)]. The major aim of the investigations is to identify genes which trigger the pollen embryogenesis species-specific and insert such a trigger-gene into the microspore. This trigger-gene should be expressed at a higher level (or lower level) to maintain the microspores to run the androgenic development instead of the regular pollen development. With the insertion of an appropriate gene the induced stress could be omitted and ideally the number of microspores, which develop into embryo-like structures, could be increased.

## **2.5 RNA-Seq Technology**

The transcriptome can be defined as the complete set of transcripts in a cell at a specific time representing a specific developmental stage or a particular physiological condition. For the profiling of the transcriptome of any organism there is a novel high-throughput RNA sequencing method which is called RNA-Seq. RNA-Seq is the abbreviation for RNA-Sequencing technology and is able to both map and quantify transcriptomes. Additionally it provides a far more precise measurement of gene expression levels than other technologies and therefore gives the possibility to quantify changes between gene expression levels. With RNA-Seq all species of transcripts can be catalogued including non-coding RNA, small RNA and so on. Furthermore RNA-Seq is capable to determine the structures of genes by using the read coverage [Wang, Z. et al. (2009)].

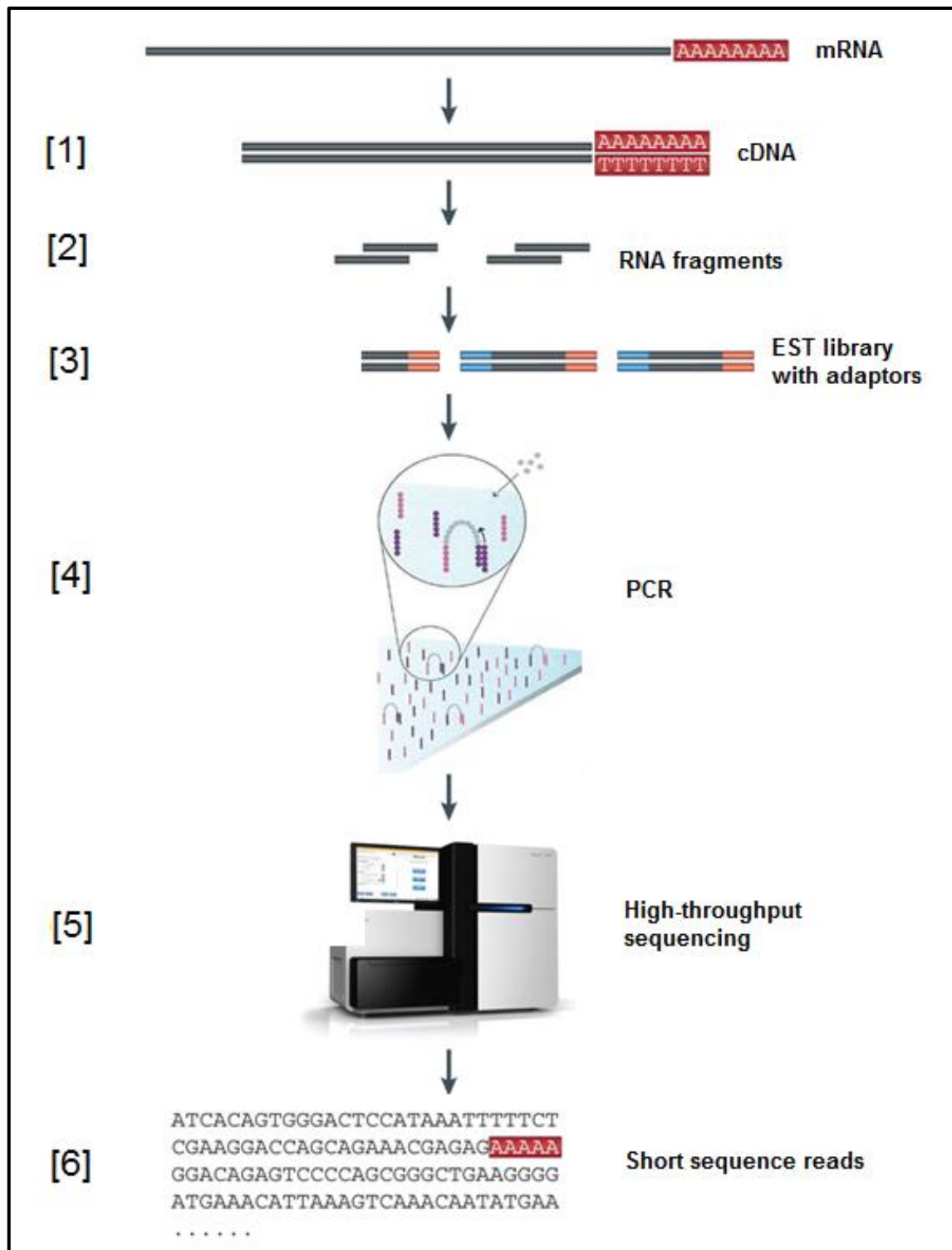
### 2.5.1 Basic Protocol of RNA-Seq

The basic protocol of the RNA-Seq technology (see Figure 3) starts with poly(A)<sup>+</sup> RNA. In the first step a double-stranded cDNA library is generated with appropriate primers [1]. After that the double-stranded cDNA sequences were fragmented into sequences with a defined length [2]. Third step is the ligation of the adapters suitable for the sequencing platform [3]. Following, the adapter-ligated cDNA fragments were amplified by PCR [4] and afterwards sequenced in a high-throughput manner [5] [Nagalakshmi, U.; Waern, K.; Snyder, M. (2010)]. The sequencing step is providing short sequence reads [6] which become assembled into RNA-Seq-Contigs *in silico*. A contig is the short term for contiguous sequence. And in bioinformatics a contig is defined as a continuous region from a set of overlapping sequence fragments of DNA or RNA (reads) resulting from their assembling.

### 2.5.2 Benefits and Challenges of RNA-Seq

RNA-Seq is especially attractive for investigations with non-model organisms because there is no need of a whole genome reference. Additionally RNA-Seq has less background sounds in comparison to microarrays for example. Moreover RNA-Seq offers an increased resolution and high rates of reproduction of technical and biological copies, which are obvious advantages of the technique not only for non-model organisms.

With the RNA-Seq technology the boundaries of transcripts could be identified very precisely. Furthermore the different lengths of reads provide variable information. Short reads deliver information about how two exons are connected. And longer reads or paired-end short reads reveal connectivity between multiple exons. Beside this, sequence variations like SNPs within the transcribed regions can be detected. The advantages of the RNA-Seq technology are the large dynamic range, the high accuracy, the high level of reproducibility and the much less requirement of RNA sample [Wang, Z. et al. (2009)].



**Figure 3 - RNA-Seq Basic Protocol [edited after Wang, Z. et al. (2009); URL-10/-11]**

The basic protocol of RNA-Seq starts with poly(A)<sup>+</sup> RNA. [1] A double-stranded cDNA library is generated. [2] The double-stranded cDNA sequences were fragmented. [3] Ligation of suitable adapters for the sequencing platform. [4] cDNA fragments were amplified by PCR (Polymerase Chain Reaction) or by bridge amplification, dependent to the sequencing strategy. [5] The fragments were sequenced in a high-throughput manner. [6] Providing of short sequence reads which become assembled into RNA-Seq-Contigs *in silico*.

But RNA-Seq also has to deal with some challenges. For example, large RNA molecules needed to be fragmented into smaller pieces, because the most deep-sequencing technologies are only capable for sequencing fragment lengths between 100 - 300 bp. Another challenge is that the reads have to be assembled into contigs without an alignment to a genomic reference sequence more often. Also sequencing errors and polymorphisms bring out some mapping problems. Only SNPs do not serve any mapping problems, because they are easily detectable polymorphisms. But in contrast, especially the uncovering of larger distances and differences often requires a genome reference or a deeper sequencing coverage, though deeper coverage requires a more sensitive sequencing depth. Finally one problem is that the larger the genome the more complex the transcriptome and the more sequencing depth is required for achieving an adequate result [Wang, Z. et al. (2009)]. Especially for *de novo* assemblies, the discovery of novel transcripts or the quantification of already known isoforms a higher depth of sequencing would be advantageous.

### **2.5.3 The RNA-Seq Data Set**

The RNA-Seq data set for this thesis was produced previously by the Illumina HiSeq sequencer. Illumina is a polymerase-based sequencing-by-synthesis method which uses bridge amplification. It uses paired-end RNA-Seq reads with separation lengths of 200-500 bp and read lengths between 35-150 bp [URL-2]. For the existing RNA-Seq data set the read lengths is 100 bp.

As initial point three different stages of barley microspores, which seemed to be important for the investigations of pollen embryogenesis in barley, were detected. The first stage (1+2) is represented by pre-mitotic microspores, which are still in regular pollen development. Second stage (3+4) is formed by embryogenesis competent microspores, which already have been stressed with the inductive treatment. And the third stage (5+6) is constituted as embryogenic pollen, one day after the treatment.

The total RNA of these microspores (and pollen cells) was extracted respectively to the previous defined stages. The amount of the RNA was analyzed with the Illumina HiSeq sequencer based on the basic RNA-Seq protocol mentioned above (see Figure 3). After the Illumina run the short paired-end RNA-Seq reads were assembled *de novo* with the program CLC. And finally the assembled RNA-Seq-Contigs were assigned to the different stages of pollen embryogenesis in barley.

Compared with the previous internship report, in this present Bachelor thesis the RNA-Seq data set is more a reference data set for the analysis steps, then a basis data set. At first the RNA-Seq data set serves as a reference to give a statement about whether a defined genomic part (Morex WGS-Contig) is expressed, in one of the previous defined microspore stages, or not. Later the data set is used for a gene prediction of the corresponding genomic region. And at last the RNA-Seq reads were used for the statistical evaluation for providing an expression pattern of the gene while the transition from microsporogenesis to pollen embryogenesis.



### **3 Methods**

#### **3.1 Processing of the Data Set**

In contrast to the internship report, the starting point for this Bachelor thesis was a previously arranged list of candidate genes. This list contains a number of genes from different organisms. These genes were selected because they were mentioned in the literature from previous investigations as initiators or participants in (pollen) embryogenesis. In this initial list there are information like the gene-symbol, the full gene name, a predicted barley transcript, an expression value and a corresponding genomic contig with associated BLASTn results assigned to every candidate gene. These genes were termed 'candidate genes' because it can be assumed that, if there is an adequate gene in the barley genome, which is expressed in the embryogenesis-competent microspore (second stage), this may be involved in the switch from regular pollen development to pollen embryogenesis. At first a detailed table (see Appendix 1) with the gene-symbol, the full name, a full description and available literature references was generated, for all candidate genes. This was made to present an overview about the already available information from existing database entries.

Out of the list containing all candidate genes a subset, due to the given expression value and the BLASTn results, was selected. This previously selected subset consists of twenty genes which could may be possible embryogenesis-trigger in barley and additionally show promising expression values in the microspores. The subset of significant candidate genes was marked in the table and their corresponding complementary DNA (cDNA) sequences were extracted for the further analysis steps.

#### **3.2 Analysis Steps**

The applications and procedures used for this Bachelor thesis consists of already used methods from the internship and some new applications suitable for the differing purpose. All performed analysis steps has been again presented in an UML activity

diagram (see Figure 4) to give an overview about used methods and intermediate steps. An activity diagram is a specific state diagram to illustrate the flow possibilities of a system. To summarize all activities, that have been done, and states, that were used to process the task settings, an activity diagram was drafted in form of a pipeline. The activities are defined as states with internal action and are represented as a single step in the flow. 'Activities' are pictured as dark green rectangles (see Figure 4). In contrast the states without internal action are drawn as light green rectangles and are termed 'objects'. The transitions between the states and activities are illustrated with arrows. All continuous arrows are transitions between two activities and termed 'Control Flows'. The dashed arrows are transitions between activities and objects, are termed 'Object Flows' and have different meanings accordant to their direction. If a dashed arrow runs from an activity to an object, the object is a state which results from the activity. But if it runs from an object to an activity, then the object is the starting state of the activity and it is required for the activity [Oestereich, B. (1999)].

In the present pipeline (see Figure 4) there are two starting points indicating two different approaches to process the task settings. The whole pipeline is divided into five parts, whereof the first and second part represents the two advances. The first approach (see [I] in Figure 4) was previously arranged and indicates the starting point by which an elaborate list of candidate genes represents an educated guess (see section 3.1). The second approach (see [II] in Figure 4) starts with the three libraries of RNA-Seq-Contigs from the three microspore stages. But this approach was not executed in the present Bachelor thesis. It needed to be mentioned because both approaches are the possible, logical entries for analyzing the transition to pollen embryogenesis in barley. The light gray background indicates that these parts were not directly performed for this bachelor thesis, even though belonging to the whole workflow.

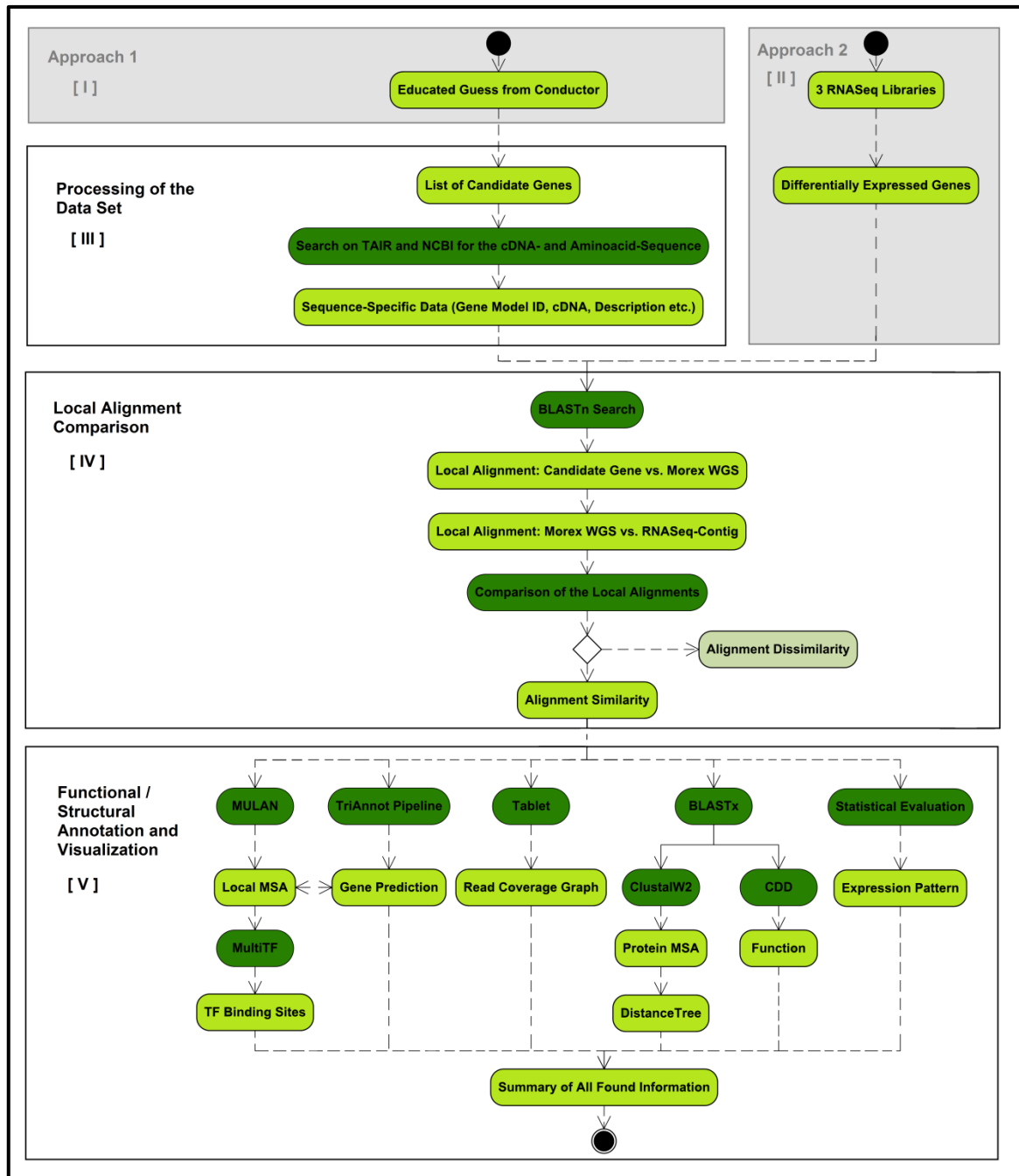
At the third part of the pipeline (see [III] in Figure 4) the processing of the Bachelor thesis is getting started. The list of candidate genes is the previously composed list

mentioned in section 3.1 as the subset. For this selected candidates the cDNA and amino acid sequences were extracted, from both public databases and internal resources from the IPK. Next step was to sum up all the available data in appropriate tables and get a first idea about the genes which have to be analyzed.

Following, the first BLASTn run [URL-4] was made against the WGS of the Morex cultivar to find associated genomic sequences (Morex WGS-Contigs) in barley to the candidate genes (see [IV] in Figure 4). And subsequently a second BLASTn search was performed to find whether the WGS-Contig was expressed in one of the microspore stages. In this step the RNA-Seq data set was used as a reference. Afterwards the local alignments of both BLASTn searches were compared, based on the start and end positions. Due to this comparison a decision was made whether the local alignments are similar or dissimilar to a certain degree. This is because if the genomic region, aligned to the candidate gene sequence, is not expressed in one of the microspore stages there is no need to further analyze this particular candidate gene. Because then there seem to be no appropriate putative homolog in barley which may be involved in pollen embryogenesis.

All candidate genes with dissimilar alignment results were omitted. And all candidate genes with promising local alignments were used for further analysis, according to the pipeline in Figure 4.

To visualize the local alignment similarity the application Mulan [URL-18] was used first (see [V] in Figure 4). Simultaneously the corresponding Morex WGS-Contig was uploaded to the TriAnnot Pipeline [URL-17] to automatically annotate the particular part with existing database resources. The local multiple-sequence alignment of Mulan and the gene prediction results of TriAnnot were compared, to confirm the supposed sequential relation. Subsequently the MultiTF tool of the Mulan pipeline was used to identify all transcription factor binding sites (TFBS) of plants in the query sequences.



**Figure 4 - Activity Diagram of the Analysis Steps**

The activity diagram illustrates the workflow (pipeline) in which the data was processed respectively to the task settings. The dark green rectangles represent activities. The light green rectangles represent the states (objects). The continuous arrows are control flows and the dashed arrows are object flows. The pipeline can be divided into five parts. [I] Approach 1: was processed previously. [II] Approach 2: is another possible advance to process the task settings. [The gray background displays that these parts were not executed while the present thesis.] [III] Processing of the Data Set: is the first part processed for the present Bachelor thesis. [IV] Local Alignment Comparison: this part again belongs to the approach 1 but also can be used for the approach 2, with only a few deviations. [V] Functional / Structural Annotation and Visualization: annotation of the genes respectively for the first approach, but it is extendable due to the task settings. For further description see the text above.

After that the mapping of the reads from the RNA-Seq data set against the WGS assembly were visualized with Tablet, to compare the expression pattern with the predicted exon-intron structure from TriAnnot. The assembly in Tablet was displayed with the coverage graph to identify the real expression of the gene from the Morex WGS-Contig in the microspore stages.

Afterwards a BLASTx [URL-5] search was made with the corresponding RNA-Seq-Contig to detect protein sequences from the different related organisms. These protein sequences were extracted to produce a multiple sequence alignment (MSA) and an averaged distance tree with ClustalW2 [URL-6]. At once the link from the BLASTx result to the Conserved Domain Database (CDD) [URL-8] was used to provide information for the functional annotation of the gene in barley.

Last step was a statistical evaluation for achieving an expression pattern of the previous annotated gene due to its expression in the three microspore stages.

### **3.3 BLAST**

BLAST is the abbreviation for Basic Local Alignment Search Tool and is one of the most used tools in computational biology. This similarity search tool was first introduced from the NCBI in 1989. The basic algorithm is very fast, statistical reliable and adaptable to variable types of sequences. This is why BLAST entrenched in the bioinformatics over the years [Korf, I. et al. (2003)].

Meanwhile there are different types of BLAST applications which cover many purposes of the scientists. First intention is to find homologous sequences and identify species. Second aim is to locate functional domains, which can be achieved by working with protein sequences. The third intention is to establish phylogeny between species on sequence scaffolds. But BLAST can only provide a first presumption about phylogenetic relations. Second to last DNA mapping is an important function, which provide the comparison of query sequences with physical chromosomal positions and it can be

used to find unknown locations of genes and functional sites. At last the algorithm can be used to map annotations from a well-known organism to an unknown [URL-3]. In this Bachelor thesis the BLAST searches were executed to find homologous regions, for achieving a presumption about possible orthologous genes, and to locate functional domains, to compare the functional structure of the query sequences.

To run a BLAST search it is possible to use the Web BLAST on the website of the NCBI or run a stand-alone command line application. For both there are some parameters which have to be defined according to the query sequence and the purpose of the search. The sequences of interest (query) and the database or reference to search in have to be determined. In command line applications also the format of the output has to be declared. Due to the purpose of the search sometimes the parameters word-size, gap-open-penalty, gap-extend-penalty, mismatch-penalty, e-value etc. have to be changed [URL-3].

### **3.3.1 BLASTn**

As first the traditional BLASTn application was used to map the cDNA sequences, of the previously selected candidate genes, onto the barley Morex WGS assembly. These analyses were performed with the IPK Barley BLAST Server [URL-4]. Based on the results from the initial list of candidate genes the most likely corresponding genomic contig (Morex WGS-Contig) was chosen. And the local alignment(s) of the candidate gene with the Morex WGS-Contig(s) was observed for the start and end position(s). The most likely corresponding Morex WGS-Contig needed to be selected manually because according to experience the hit with the highest score must not be the correct WGS-Contig. To find out which is the corresponding contig the score, the percentage identity and the e-Value were compared. In fact the parameters have to be compared and interpreted for each case individually. Therefore the selection of the correct Morex WGS-Contig depends on experience and practical knowledge.

With the Morex WGS-Contig, as initial point, a second BLASTn search was made against the set of RNA-Seq-Contigs representing the transcriptomes of the three microspore stages. For these analyses the version BLASTn 2.2.25+ was used as a command line application with default settings. Subsequently the two local alignments were compared by the start and end positions, to determine whether the region of the genome, similar to the candidate gene, is expressed in the microspores of barley or not. Due to this, a statement about the alignment similarity or dissimilarity was made.

### 3.3.2 BLASTx

Later a BLASTx search was executed with the RNA-Seq-Contig to find homologous proteins and to confirm the association between the initial candidate gene and the expressed barley equivalent. The BLASTx algorithm translates a nucleotide query sequence into the 6-frame translation products and searches, with this translated sequences, against given protein subject sequences or a protein database [URL-3].

The BLASTx application was used with default settings performed with the program of the NCBI [URL-5]. The intention was to achieve similar protein sequences from some organisms related to barley. These organisms are purple false brome (*Brachypodium distachyon*), millet (*Sorghum bicolor*), rice (*Oryza sativa*), maize (*Zea mays*), bread wheat (*Triticum aestivum*), tobacco (*Nicotiana tabacum*) and thale cress (*Arabidopsis thaliana*). Those introduced organisms already were in focus while the analysis from the previous internship report. At this point the link to the CDD, providing a scheme about domains matching onto specific regions of the query sequence, was reused.

### 3.4 ClustalW2

ClustalW2 produces biological significant multiple sequence alignments (MSA) for a set of DNA sequences or protein sequences [URL-7]. It calculates the most likely MSA in three basic steps. As the first step all sequences of the input set were aligned pairwise (global) and arranged due to their alignment score. Consequential a distance matrix is

formed with the calculated arrangement in the second step. Finally a MSA is produced out of the distance matrix.

The main intention in generating a MSA is to identify conserved sequence regions, sites or domains in the input sequence set. This can be achieved by including known annotated sequences into the analyses. The second general aim is to show evolutionary relationships and shared lineages between various species. The phylogenetic relation can be shown with the Cladograms or Phylograms ClustalW2 produces automatically beside the MSA [URL-6; URL-7].

The protein sequences of the previous specified organisms (section 3.3.2) found with the BLASTx search were combined in a file and uploaded as the input for the ClustalW2 application. These protein sequences are expected to be homologous. Therefore the gap-open penalty was set to 100 and the gap-extension penalty was set to 10.0 in the pairwise alignment options as well as in the multiple sequence alignment options. The gap penalties were set to maximum because high penalties produce compact sequence alignments. ClustalW2 just produces global alignments but the protein sequences that needed to be aligned are from different organisms and thus it may be supposed that only local regions have striking similarities. These local similar regions can only be detected in a global alignment through maximizing the cost for creating a gap and for extending the gap.

After the calculations were finished the JalView-application was used to visualize the MSAs most suitable. In the menu bar of the JalView-tool the color for the MSA was set to 'Percentage Identity'. This highlights the amino acids with a color scale from blue tones, according to the number of occurrence in the MSA for each position. Additionally the 'calculate tree' option was used from the menu bar, to open the 'Average Distance Tree Using % Identity' in a second window. This distance tree displays the evolutionary relation between the input sequences of the MSA. This graphical presentation of a distance tree with JalView should provide a statement



about the phylogenetic and functional relationship of the analyzed sequences on amino acid level.

### 3.5 Conserved Domain Database

The Conserved Domain Database (CDD) is a protein annotation resource that consists of well-annotated multiple sequence alignment models for protein domains and full-length proteins [URL-8]. Especially the database is representing protein domain models that are conserved in molecular evolution [Marchler-Bauer, A. et al. (2009)]. The CDD mainly consists of domains curated by the NCBI. These domains are annotated with 3D structures, defined domain boundaries and an insight into sequence-structure-function relationship if present. Further content of the database are domain models from external source databases like Pfam, SMART, COG or TIGRFAM. The main properties of the database are to visualize the architecture of protein domains, highlight the presence of domains on the sequence, identify a putative function of an unknown protein sequence and eventually identify specific amino acids in a sequence that are putatively involved in functions as DNA binding or catalysis. Especially conserved domains are in the center of attention and for that reason information about their distinct function, structural units, building blocks and conserved patterns or motifs were summarized for scientific benefits [URL-8]. In the graphical output the domain annotation is displayed as specific domain matches, non-specific hit, conserved domain superfamily and functional sites according to the query sequence [Marchler-Bauer, A. et al. (2009)]. Due to the use of the Reverse Position Specific (RPS) -BLAST algorithm conserved domains can be identified very fast for an unknown protein sequence. The RPS-BLAST searches with the query sequence against a database of profiles, which are domains in this case. Because of the linking to the BLASTx application, the CDD takes the translated query sequence directly and compares it with the curated domain set and all domain models from external databases. Therefore the CDD provides information about the topology and the functional role of all domains which have matched on the query sequence [URL-8]. All gathered information regarding the domains for the RNA-Seq-Contigs, with the help of the CDD, were processed manually

and arranged in appropriate formats. Therefore screen shots were made of the results presented on the website of the CDD and the best hits were extracted from these pictures if required.

### 3.6 Mulan

Mulan is the abbreviation for multiple-sequence local alignment and is a new integrative comparative application that generates textual and particularly graphical multiple-sequence local alignments (MSLAs). The tool Mulan produces rapid, dynamical and very accurate local alignments for both closely and distantly related organisms. Especially for distant organisms Mulan ensures a reliable representation of short- and large-scale genomic rearrangements. For the graphical visualization of the finished MSLAs there are different options for achieving a suitable presentation, e.g. the reference sequence of the pairwise alignments can be changed flexibly.

Due to the fact that regions conserved in various species often correlate with functional elements, Mulan is capable to identify those conserved elements specifically in an evolutionary context. Several data analysis and visualization schemes within the interface of Mulan allow the identification of coding and non-coding elements as well as sequence arrangements like inversions, transpositions and subsequence reshuffling. These conserved elements provide information about the complexity of evolutionary sequence movements or variation even across large distances.

The tool Mulan consists of three interrelating algorithms: TBA, MultiTF and GALA. The Threaded Blockset Aligner (TBA) is a program which produces blockset alignments with selectable reference sequences. These alignments are a suitable presentation of local pairwise and multiple-sequence alignments. Additionally the TBA program detects those evolutionary conserved elements within the alignment, described above. The second algorithm is the MultiTF program which is implemented to identify evolutionary conserved transcription factor binding sites (TFBS) within all the input sequences in the MSLA. The last mentionable application is the two way communication with the GALA, which is the abbreviation for Genome Alignment and Annotation Database.

Summarized Mulan can be used to (a) produce graphical and textual local alignments, (b) determine phylogenetic relationships, (c) generate phylogenetic trees and (d) detect evolutionary conserved regions (ECR). For those ECRs Mulan has an additional module termed ECR-Browser, filtering for conserved regions. This may be important for generating hypotheses about the function of the mentioned regions [Ovcharenko, I. et al. (2005)].

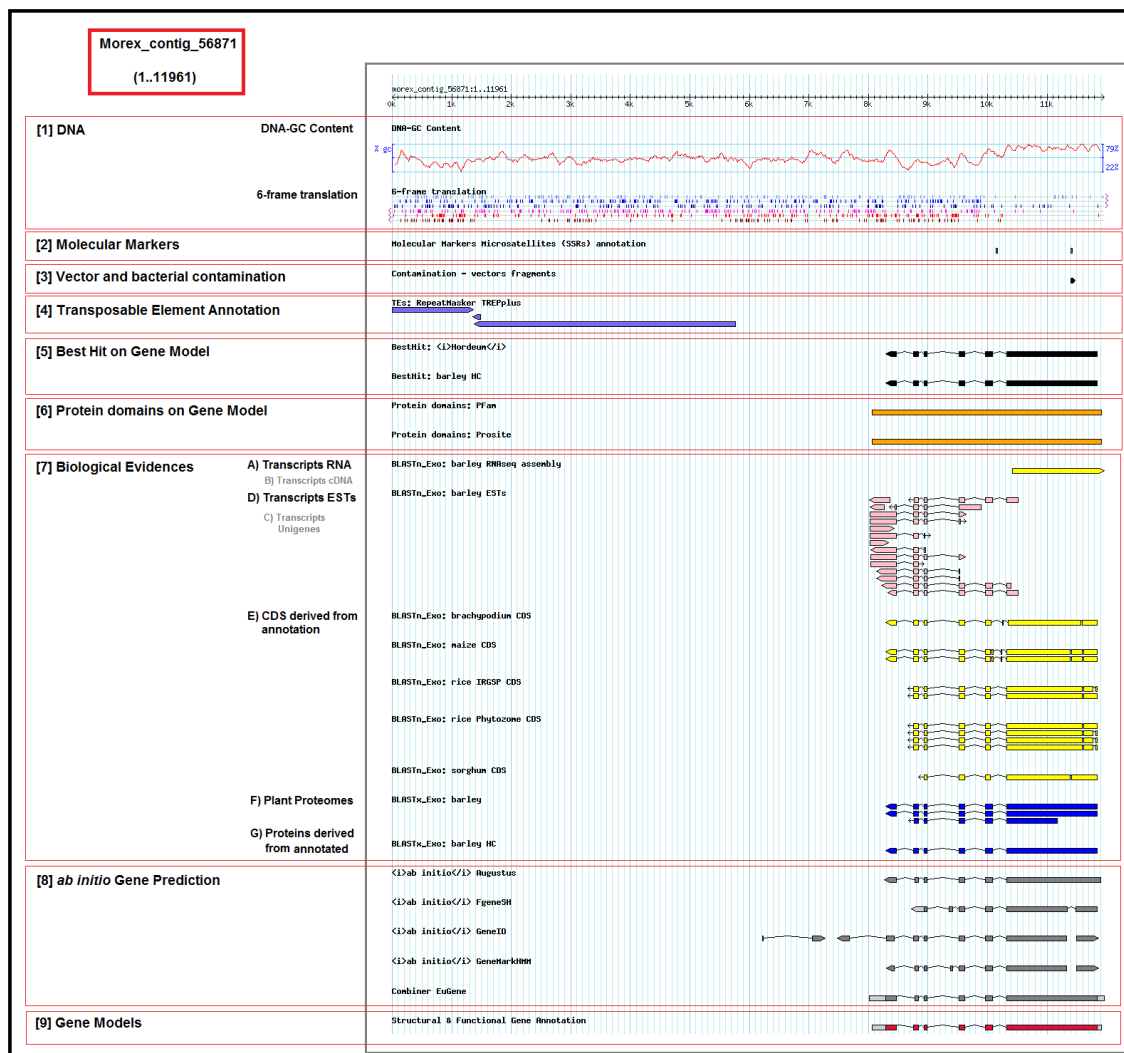
### 3.7 TriAnnot Pipeline

TriAnnot is a pipeline for the automated structural and functional annotation of plant genomes, which can be accessed through a web interface for small scale analysis. The TriAnnot pipeline was developed to be able to deal with large and complex genomes, because this becomes particularly with the advent of the high-throughput NGS technologies. The pipeline has a modular architecture allowing a simultaneous annotation of protein-coding genes, identification of conserved non-coding sequences and detection of molecular markers. TriAnnot combines methods and applications from other pipelines with the intention to integrate the most innovative features of these already available pipelines [Leroy, P. et al. (2012)].

To start a small scale analyses the user can submit query sequences in FASTA format with a size between 10 Kb and 3 Mb. If the analysis was processed successfully the results can be displayed with GBrowse. GBrowse is a conjunction of databases and interactive web pages used for the manipulation and the displaying of genome annotation. The TriAnnot pipeline provides a range of tracks that can be selected individually by the user to view in the browser (see Figure 5).

The selectable tracks are: [1] DNA (6-frame translation, GC content); [2] molecular markers; [3] vector and bacterial contaminations; [4] transposable element annotation; [5] best hit on gene model; [6] protein domains on gene model; [7] biological evidences (A-G); [8] *ab initio* gene prediction and [9] gene models. For the present analyses the genomic WGS-Contig was uploaded in FASTA format and all tracks that show significant interesting results were selected individually for the

respective genomic sequence. As a suitable output the results were exported into PNG-files [URL-17].



**Figure 5 - GBrowse Viewer of TriAnnot**

GBrowse screen of the TriAnnot pipeline, with all available features calculated for this especially Morex WGS-Contig. The genomic contig is represented at the whole width of the figure. The colored stripes are matching outputs from different databases and sources. For further description see the text above.

### 3.8 Tablet

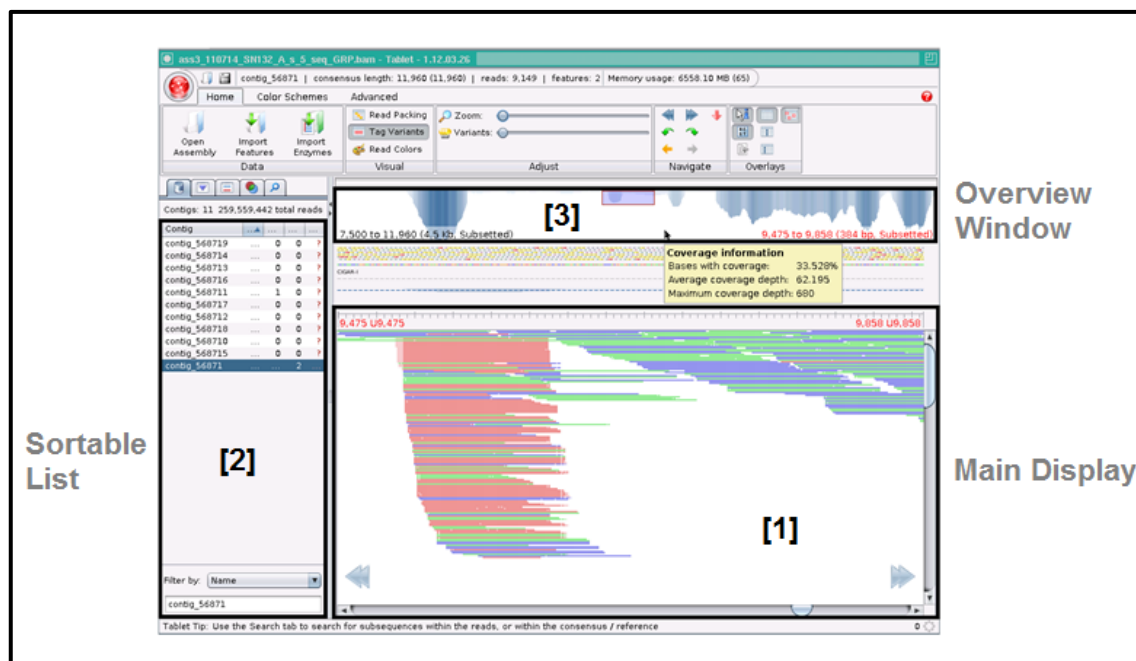
Tablet is a software for the visualization of next-generation sequence assemblies and alignments, which is freely available and employable for users of all abilities. The software was developed at The James Hutton Institute and was published in December 2009 first [URL-9]. Tablet supports most of the common input assembly formats and is

able to process a large number of reads. The interface of Tablet provides an overview of the whole selected contig and therefore gives access to any region of the assembly with an intuitive navigation. All regions of the assemblies are displayed with high-quality at any zoom level.

The interface of Tablet can be divided into several areas. The main display (see [1] in Figure 6) shows all reads aligned against the selected contig running from left-to-right across the screen. These reads can be displayed at scaling zoom levels. This display should draw attention to informative regions at first glance. Supportingly the reads can be colored in different schemes to point out these informative regions. The default coloring is the basic nucleotide scheme by which every base is assigned to another color. With this graphical color scheme sequence-composition pattern like microsatellites, poly-A-tails, mononucleotide runs or GC rich regions can be detected. Further color schemes indicate the direction of the reads, the lengths of the reads or rather the read type. Every scheme support varying functionalities, but all schemes provide the visualization of SNPs or sequencing errors, as they highlight varying bases. On left hand of the interface there is a sortable list (see [2] in Figure 6) of all available contigs that could be visualized in the main display [1] and therefore represent the reference sequences. These contigs can be sorted due to their name, their contig length, and the number of reads or their previous defined features.

Above the main display there also is an overview window (see [3] in Figure 6). This displays a scaled-to-fit summary or a coverage graph of all the reads mapped to the selected contig. The length of the window represents the length of the contig independent of the zoom level of the main display.

Between this overview window and the main display all six reading frames of protein translation can be optionally showed. Detailed information about the coverage can be seen with the mouse-over function in the overview window [Milne, I. et al. (2010); Milne, I. et al. (2012)].



**Figure 6 - Interface of Tablet**

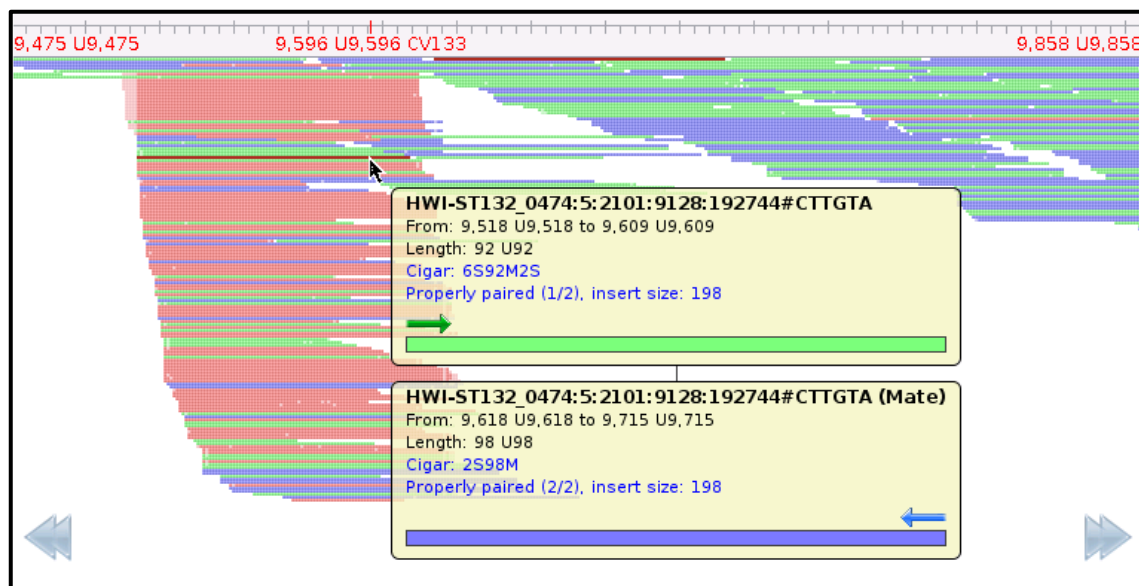
The GUI of Tablet can be divided into the main display (1), a sortable list (2) and the overview window (3). The main display contains the assembly of the reads mapped to the selected contig in varying zoom levels. The list on left hand contains all contigs of the assembly. This list can be sorted due to the features of the contigs. These are the contig name, the contig length, the number of reads or a previous defined feature based on the assembly. The overview window shows either an scaled-to-fit summary or a coverage graph of all reads aligned against the selected contig.

Additionally there is also an existing mouse-over function for the main display which provides detailed information about the single reads. The additive window contains the read name, its location, length and orientation. Also there can be seen either the position of the paired read (see Figure 7) (and the insert size) or that the mate is unmapped (see Figure 8).

Mate-pair or paired-end sequencing provides many advantages especially for *de novo* transcriptome sequencing approaches. Small insert sizes between paired reads provide precise resolutions of non-repetitive sequence regions. In contrast long insert sizes can help to rearrange or untangle contigs over a large mapping. Additionally contigs can be orientated or located in a scaffold. Scaffolds are formed by contigs separated by gaps of known lengths.

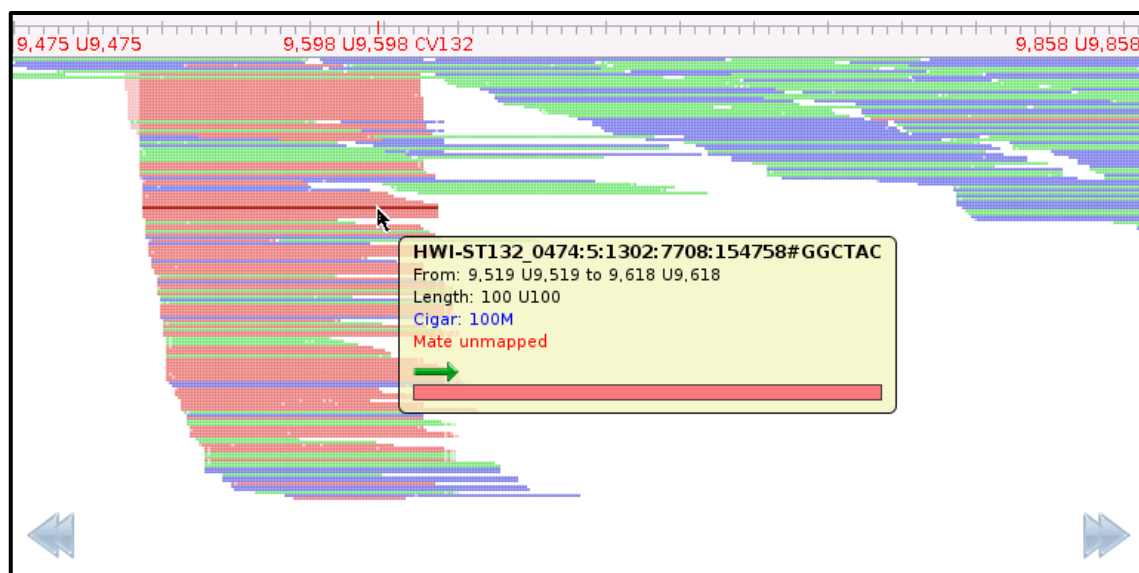
In Tablet Illumina mate-pair reads mostly map to the same reference sequence caused by the small insert size and which is helpful for the coverage graph interpretation.

Reads whose mate is unmapped are exceptional cases and may indicate that the read is mapped false positive at this position.



**Figure 7 - Main Display of Tablet Showing Properly Paired Reads**

Part of the read coverage graph of a Morex WGS-Contig. The mouse-over function of Tablet is showing particular information about the selected read and its corresponding mate-pair read. The green reads are the forward reads mapped to the reference sequence and the blue reads are the associable mate-pair reads.



**Figure 8 - Main Display of Tablet Showing a Read with Unmapped Mate**

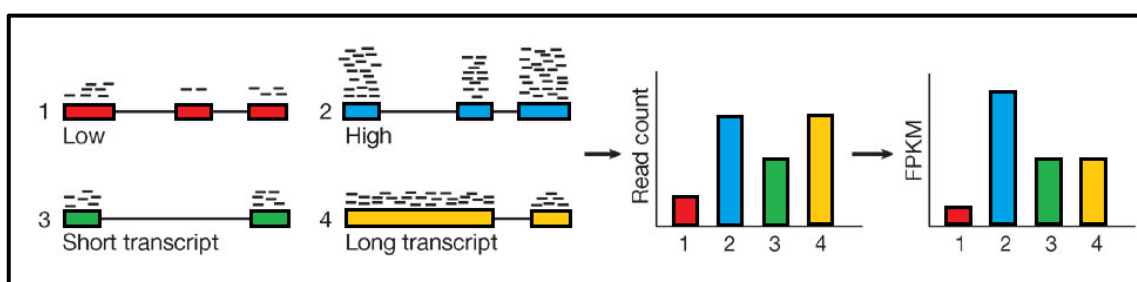
Part of the read coverage graph of a Morex WGS-Contig. The mouse-over function of Tablet is showing particular information about the selected read. The red reads are reads whose mate-pair reads are unmapped to the particular reference sequence.

### 3.9 Statistical Evaluation

After the detailed annotation of the genomic sequence (Morex WGS-Contig) the expression level of the transcript (RNA-Seq-Contig) within the three microspore stages is important for further molecular biological applications.

Starting with the reads produced by the RNA-Seq technology the transcripts were assembled into RNA-Seq-Contigs using the CLC program. Following the reads were re-mapped against the assembled RNA-Seq-Contigs to calculate the coverage. This coverage is a value which represents an averaged number of reads per nucleotide and is varying due to the different stages of microspore development. This number of reads or coverage is also termed read count.

Caused by some variability occurring with RNA-Seq technology the read count needed to be normalized for a comparable evaluation (see Figure 9). Variability is produced by the RNA fragmentation during the cDNA library construction (see (2) in Figure 3) and also is caused by the differing number of reads, each run. An appropriate program which properly normalizes the RNA-Seq read counts into FPKM values is Cufflinks. It helps estimating and comparing transcript expression levels. The abbreviation FPKM stands for 'Fragments Per Kilobase of exon per Million fragments mapped' and is used for paired-end sequencing reads.



**Figure 9 - Read Count Normalization [edited after Garber, M. et al. (2011)]**

Due to the number of reads (1 and 2) and different cases of transcripts (3 and 4) the read count (read coverage) has to be normalized for a comparable evaluation. In the left there are the four different variants of transcripts that can be detected by RNA-Seq technology. In the middle there is a graph displaying the number of reads that were detected for the four transcripts. Rightmost the graph shows the read count after it was normalized into FPKM values.

For every RNA-Seq-Contig the read counts were calculated in association to every stage of microspore development (see Appendix 2). And subsequently the expression



pattern of an RNA-Seq-Contig across the three stages was visualized by a bar chart produced with Microsoft Excel. The normalization was not performed because the number of reads from the three stages was comparable. Moreover there was no time within the framework of the Bachelor thesis to realize the normalization step.

**Table 1 - Summary of the Methods**

A Summary of all the methods and used tools while the present Bachelor thesis. First column contains the used tools. Second column contains the required input. Third column contains the output, produced from the application. -q: Query; -db: Database; -o: Output; -outfmt: Outputformat; s: Subject; bp: base pairs; Mbp: Mega base pairs.

Tool	Input	Output
BLASTn	<p><u>Query:</u> (-q) <b>Nucleotide sequence</b> (cDNA) of a RNA-Seq-Contig in <b>Fasta format</b>.</p> <p><u>Database:</u> (-db) <b>Whole genome sequence assembly of the Morex cultivar</b> (Set of genomic nucleotide sequences).</p>	<p><u>Out:</u> (-o) <b>'blastn'-file</b> with all results listed up in the predefined output-format</p> <p><u>Output-Format:</u> (-outfmt) <b>"7 query id; subject id; query length; q. start; q. end; subject length; s. start; s. end; score; evalue; % identity; alignment length"</b></p>
BLASTx	<b>Nucleotide sequence</b> of a RNA-Seq-Contig in Fasta format.	<p><b>Similar protein sequences</b> to the translated nucleotide sequence <b>from variable</b>, but mostly, <b>related organisms</b>. Link to their entries in the NCBI database.</p> <p><b>Scheme about domains</b> which match with a specific region of the query sequence. (link to the CDD)</p>
ClustalW2	File with <b>similar protein sequences</b> from related organisms in Fasta format.	<p>Graphical presentation of a <b>multiple sequence alignment</b> with all input protein sequences in JalView application.</p> <p><b>Average distance tree</b> using the percentage identity with marked distances.</p>
CDD (Conserved Domain Database)	<b>Query protein sequence</b> from the BLASTx search.	Information about the <b>topology</b> and <b>functional role of putative domains</b> on the query.
MULAN	Selectable number of nucleotide sequences in Fasta format containing an alignment.	<p>(1) Dynamic visualization of an <b>Multiple-Sequence local alignment</b>.</p> <p>(2) <b>TFBS</b> conserved across all the species.</p>

Tool	Input	Output
TriAnnot Pipeline	File with up to 10 <b>nucleotide sequences</b> of RNA-Seq contigs in <b>Fasta format</b> . (> 1000 bp and < 3 Mbp)	<b>TriAnnot GBrowse</b> with a range of tracks that can be selected: (1) <b>gene models</b> ; (2) <b>ab initio gene prediction</b> ; (3) <b>biological evidences</b> ; (4) <b>phylogenetic evidences</b> – protein (best hit – BLASTp); (5) <b>transposable elements</b> annotation; (6) <b>conserved non coding sequences</b> ; (7) <b>vector and bacterial contaminations</b> ;(8) <b>molecular markers</b> ; (9) <b>DNA</b> (6-frame translation, GC content). Output format as low-resolution <b>PNG</b> , <b>GFF</b> annotation table or <b>FASTA</b> sequence file.
Tablet	(1) <b>Primary assembly</b> file or URL (e.g.: BAM assembly)  (2) <b>Reference/consensus</b> file or URL (e.g.: FASTA format)	<b>Vizualization of the primary assembly</b> . Main Display: <b>Selected contig with the mapped reads</b> . Overview Window: <b>Coverage Graph</b> over the whole contig length.

## **4 Results**

For this present Bachelor thesis a total of twenty candidate genes were analyzed *in silico* with the introduced methods. Furthermore a new folder was created for every gene, where all files of the important sequences and all pictures provided by the tools were stored. With the help of the files and the readme file all analysis steps can be reproduced and understood. These files and folders are located on the CD enclosed.

The abscisic acid insensitive 3 gene (ABI3) from *Arabidopsis thaliana* and the associated barley homolog will be presented in detail in this thesis. This representative example was chosen to demonstrate the interpreting of the methods and to point out the main results that can be achieved. With the help of the elaborated interpretation and evaluation of the ABI3 gene the analysis of the other candidate genes will become somewhat easier and faster.

The ABI3 gene is a transcription factor known from *Arabidopsis thaliana*. And due to the literature it is a homologous gene to the maize transcriptions factor Viviparous-1 [URL-20]. The gene contains a B3-DNA-binding domain, which is necessary for the specific interaction with a highly conserved motif, present in many seed-specific promoters. The B3 domain constitutes, in association with an activation domain, the transcriptional activity of ABI3. The gene was detected through mutations which lead to various aberrations during embryogenesis. Today ABI3 is known as a regulator of the transition between embryo maturation and early seedling development [URL-20], where it interacts with both the FUS3 and LEC1 gene. ABI3 is a seed-specific transcription factor which is crucial in maturation processes. It is also involved in auxin pathways and sideways of the root development in Arabidopsis [URL-20].

Out of the twenty analyzed genes only the agamous-like 38 gene (AGL38) could not be further analyzed, because there could no appropriate sequences in barley found that

could be assigned to the gene unambiguously. But that does not mean that there is no appropriate sequence in barley.

#### 4.1 Structural Gene Annotation

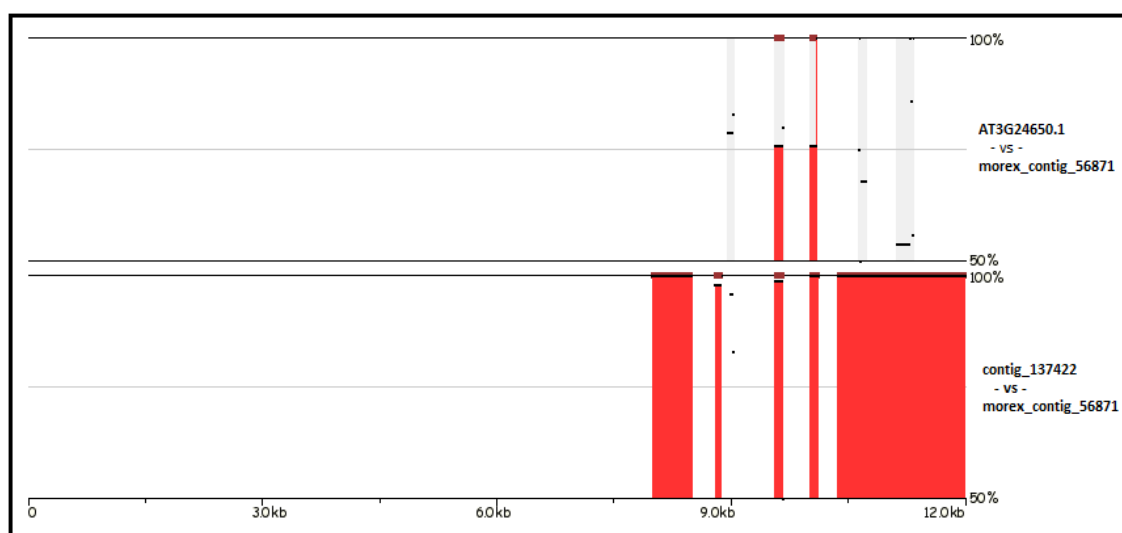
First part of the analysis was the structural gene annotation, which should provide information about structural analogies and differences between the candidate gene and the associated sequences in barley. Moreover the analysis steps should resolve the exonic structure and the transcription boundaries of the Morex WGS-Contig in barley. The structural gene annotation is simply based on sequence comparison on nucleotide level as well as on amino acid level.

Main reason for the structural annotation is that structural conserved regions experientially go with conserved function in phylogenetic context. That means if a gene in barley could be detected, which is just partly homolog to an already known gene, it is possibly involved in comparable pathways and may also has an equal function with an increased probability.

As starting point the local alignments between the candidate gene, the RNA-Seq-Contig and the genomic Morex WGS-Contig were visualized with the tool Mulan (see Figure 10). In contrast to other multiple sequence alignment (MSA) tools, Mulan presents stacked-pairwise local alignments. Which, in this case, helps to discriminate between coding and non-coding regions and also shows conserved regions. The MSLA in Figure 10 represents the pairwise alignment of the ABI3 gene (AT3G24650.1) against the associated genomic Morex WGS-Contig (morex\_contig\_56871) arranged on the top of the alignment of the RNA-Seq-Contig (contig\_137422) associated to the WGS-Contig.

The red bars in Figure 10 are intergenic elements and represent so called evolutionarily conserved regions (ECRs) with a set similarity of at least 50 %. The similarity was set to 50 % to find regions that are comparable but also detectable in sequences of different species. The shaded gray bars are alignments resulting from the

reverse strand [Ovcharenko, I. et al. (2005)]. Due to the visualization the RNA-Seq-Contig may consists of five exons of different lengths and may have two exons in common with the ABI3 gene from *Arabidopsis thaliana*. As expected the similar regions between the both sequences from barley are larger than the aligned region between the Morex WGS-Contig and the ABI3 gene from *Arabidopsis thaliana*.

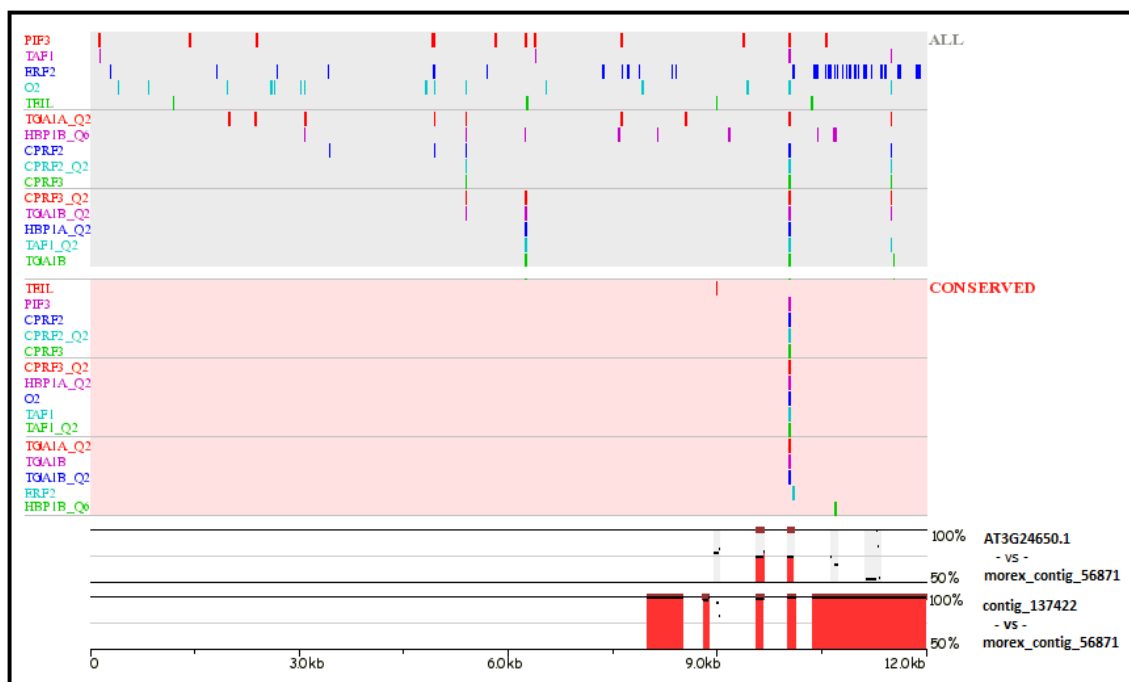


**Figure 10 - Multiple-Sequence Local Alignment of Mulan**

Stacked-pairwise alignment of the ABI3 gene (*Arabidopsis thaliana*) and the corresponding genomic sequence (morex\_contig\_56871; *Hordeum vulgare*) and RNA-Seq-Contig (contig\_137422; *Hordeum vulgare*). Red bars are intergenic elements with a set sequence similarity of at least 50%. Gray bars are alignments resulting from the reverse strand. The length of the alignment is the length of the base sequence, in this case the genomic sequence.

Additionally Mulan can submit the MSLA to the application MultiTF, which is able to identify transcription factor binding sites across the input sequences. The detection of conserved TFBS is helpful for specifying conserved functional regions on the sequences involved in transcriptional regulation. MultiTF cannot be accessed stand-alone therefore it is an extension of Mulan, GALA and the ECR Browser of the NCBI. For the present MSLA the TRANSFAC library [URL-18] of plants was chosen and all species specific transcription factors were selected for screening and afterwards the identified TFBS were visualized in another stacked profile (see Figure 11). The TRANSFAC library of plants contains all known TFBS. And therefore all available TFBS were visualized, because the screening was supposed to just provide a hint about the presence or absence of those nucleotide patterns.

In the gray part of the Figure there are all TFBS that were found within the sequences separately. The TFBS are color coded to distinguish between the position-specific hits. For the detection of TFBS there are several methods and algorithms which are able to find putative binding sites. Due to especially consensus sequences or nucleotide patterns these specific TFBS are defined through several investigations and kept in databases. The MultiTF application just uses an algorithm to find those considered nucleotide pattern within the input sequences. But those consensus sequences which form binding sites for transcription factor proteins are nucleotide sequences, which also occur regularly a number of times within the whole length DNA. Therefore the MultiTF algorithm detects the TFBS repeatedly as illustrated in the gray part of Figure 11.



**Figure 11 - Transcription Factor Binding Sites Profile of MultiTF**

Transcription factor binding sites upon the stacked-pairwise alignment of the ABI3 gene (*Arabidopsis thaliana*) and the corresponding genomic sequence (morex\_contig\_56871; *Hordeum vulgare*) and RNA-Seq-Contig (contig\_137422; *Hordeum vulgare*). The gray figure part contains all TFBS that were found in the sequences. And the red figure part contains the conserved TFBS that were found in all three sequences. The stacked-pairwise alignment presents ECRs with at least 50% sequence identity.

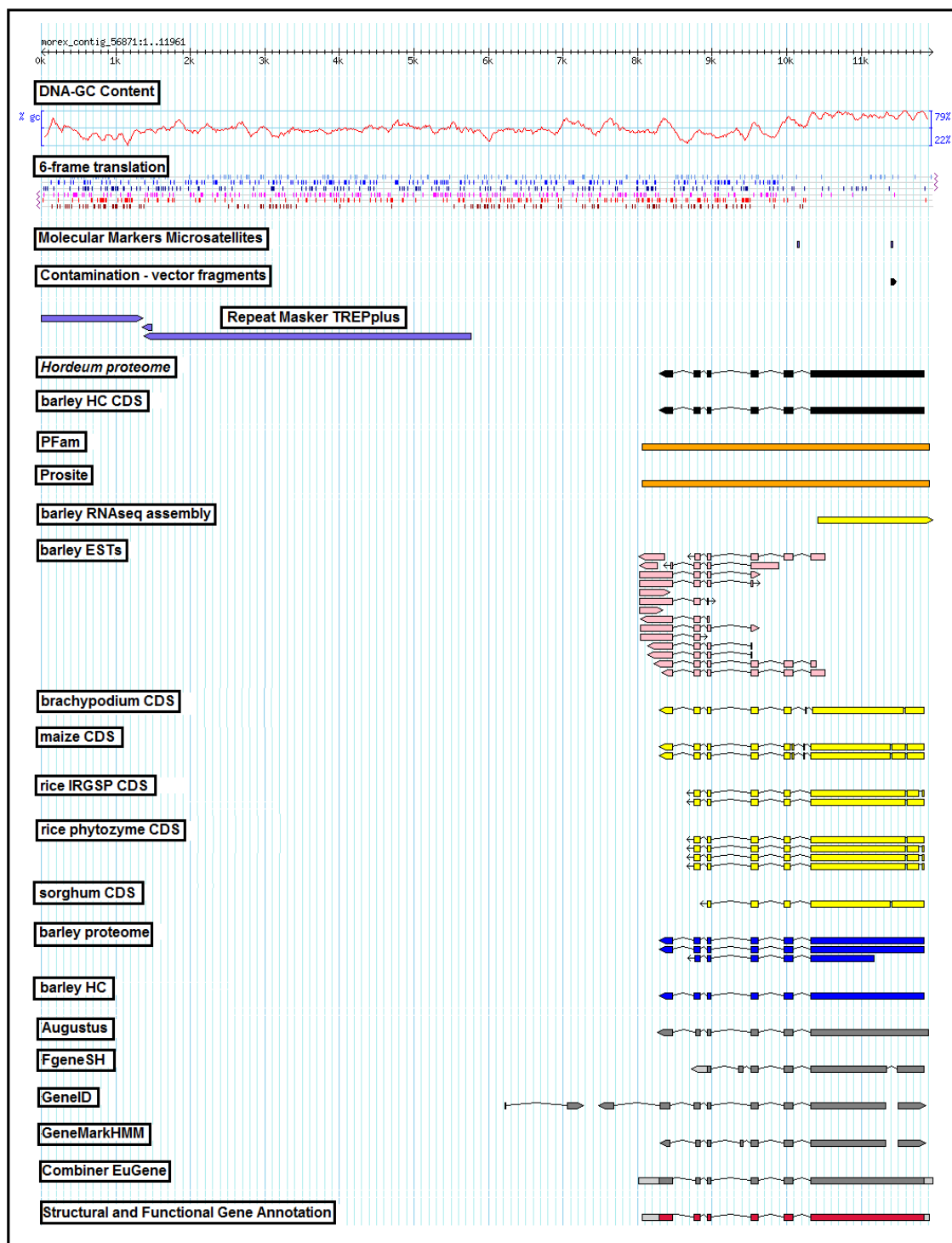
Beneath in the red part of the Figure only the evolutionary conserved TFBS were visualized which could be detected in all three sequences. These are the TFBS which

meant to be functionally sites where a transcription factor protein actually binds. And those conserved TFBS are relevant in the end.

Supplementary the genomic sequence of the Morex WGS-Contig was loaded into the TriAnnot pipeline for an independent gene prediction. This application was performed to identify the correct direction of the gene as well as the start and end positions of the gene. Furthermore the availability of homologous genes from related organisms in public databases could be observed. The intention of these automated analyses is to confirm the positions of expressed regions in the genomic Morex WGS-Contig. And additionally have a suitable adjustment against *Triticeae* databases.

Figure 12 shows that the Morex WGS-Contig divides into two parts. The first part is predicted as a repetitive region described as 'Repeat Masker TREPplus' (two purple stripes). The second part of the Morex WGS-Contig seems to be a gene with five or six exons, depending to the selected track. This part represents an expressed region and is about one third of the entire genomic sequence. The two black stripes indicate "BestHits" from public available barley gene models, according to the internal description. Beneath this, the two orange stripes show that for this position there are conserved domains, stored in the protein domain databases Pfam and Prosite. The pink stripes display available ESTs of barley which represents parts of an active gene. Moreover there are many CDS presented through yellow stripes, which derived from full-length cDNAs from different organisms belonging to the family *Poaceae*. In contrast the four blue stripes are matching proteins sourced from the barley proteome.

At last there are five gray stripes provides by five different gene prediction software approaches. Due to the different algorithms used by the software the predicted exon-intron-structures deviate from each other somewhat. In the final row there is a gene model with and functional and structural annotation achieved by all the matching references and predictions.



**Figure 12 - Morex WGS-Contig Visualization in the GBrowse Viewer of TriAnnot**

GBrowse screen of the Morex WGS-Contig 56871, with all available features calculated by the TriAnnot pipeline. The genomic contig is represented at the whole width of the figure. The colored stripes are matching outputs from different databases and sources. For further description see in the text above and Figure 5.

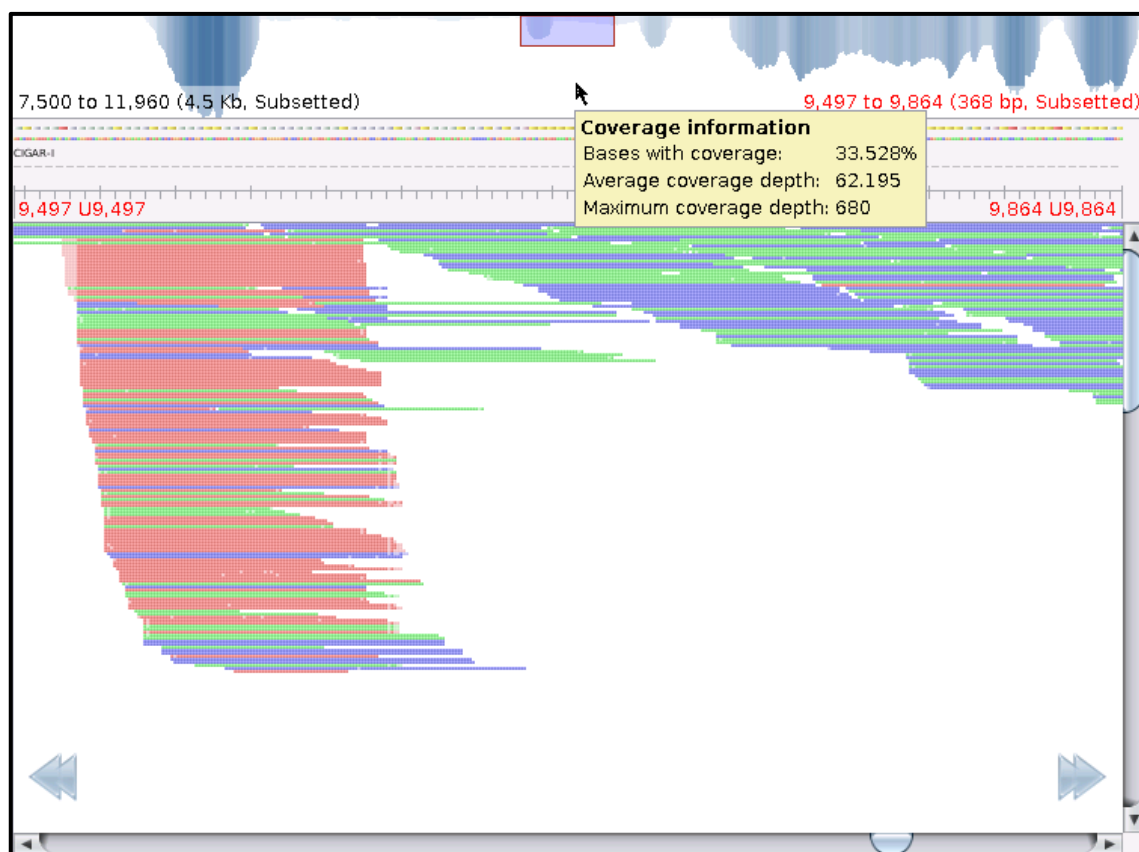


For all hits there are global information like the type of sequence (mRNA, match or repeat-region), the start position and the end position of the gene. In the GBrowse of TriAnnot there is also additional information about the coverage of the sequence, its length, the score, the position in the genomic contig and available annotation features. These global and additional details are visible with a mouse over function or a click on the feature dedicated to each available stripe.

By comparing Figure 10 of the MSLA from Mulan and the Figure 12 of the prediction from TriAnnot it can be seen that both applications independently show similar results. At the end of the Morex WGS-Contig an expressed gene can be detected in barley as well as in other related organisms. Due to the results the gene may be transcribed from right to left and may have six exons altogether, starting with a large one and followed by five small exons. The fourth exon seemed to be unrepresented in the local alignment of Mulan because of the short sequence length.

Afterwards the sequence assembly of the RNA-Seq-Reads mapped against the regarding Morex WGS-Contig was inspected using Tablet. This mapping tool provides a first overview about the number and the range of reads that could be remapped to the reference sequence. The entirety of the read coverage represents the transcribed region including eventual mapping errors. In Figure 13 the main display and the overview window of Tablet are represented. The overview window shows the part of the WGS-Contig to have a closer look. The coverage of the reads over the whole Morex WGS-Contig length amounts 33.528 %. In this case this value quantifies the percentage of the whole WGS-Contig length covered by RNA-Seq transcripts. The read coverage of 33.528 % supports the predicted expressed region in TriAnnot in ratio to the whole contig length.

Unfortunately the coverage graph in the overview window is dissimilar to the exon-intron-structure predicted previously. This is why it will be necessary to have a closer look at the reads and to all 6-frames of protein translation in tablet.



**Figure 13 - Read Coverage Graph of the Morex WGS-Contig in Tablet**

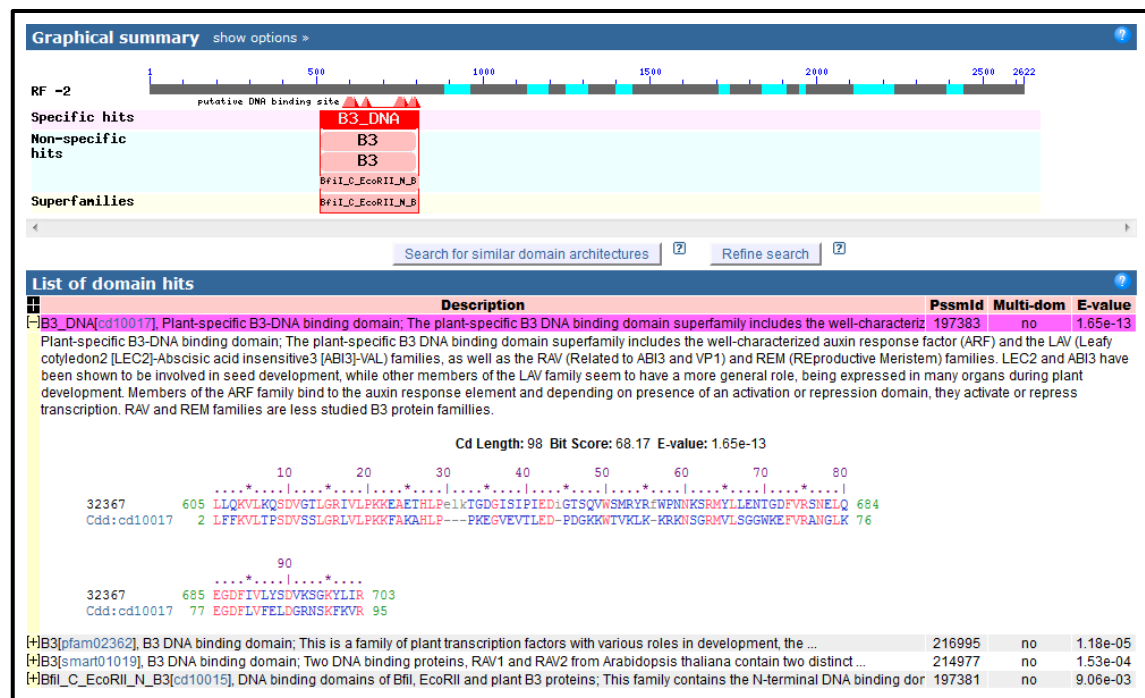
Main display and overview window of tablet showing the mapping RNASeq reads of all three microspore stages against the Morex WGS-Contig which is a pendant to the ABI3 gene. The figure displays a subset from 7500 to 11960 kb of the WGS-Contig. The coverage of the reads amounts 33.528 % across the whole WGS-Contig length. The green lines are forward reads and the blue lines are the properly paired mate reads. The red lines are reads which mate reads are unmapped on the reference sequence.

## 4.2 Functional Gene Annotation

Second part of the analysis was the functional gene annotation. This should give a suggestion about the functional role of the expressed gene in barley and the differences on amino acid level to similar proteins in related organisms. Furthermore the analysis steps may point out the (conserved) functional region in the protein. In the end the start and end position of the gene is important for further analysis with molecular biological techniques.

At first the RNA-Seq-Contig which is associated to the genomic homolog of the candidate gene was loaded as query sequence into the BLASTx application on the NCBI. The RNA-Seq-Contig was translated into a protein sequence and used as input

sequence for the similarity search. This search was processed on amino acid level because differences in protein sequences do not attach that much importance than differences on nucleotide sequences. The BLASTx run was processed to find similar or homolog proteins in related organisms.



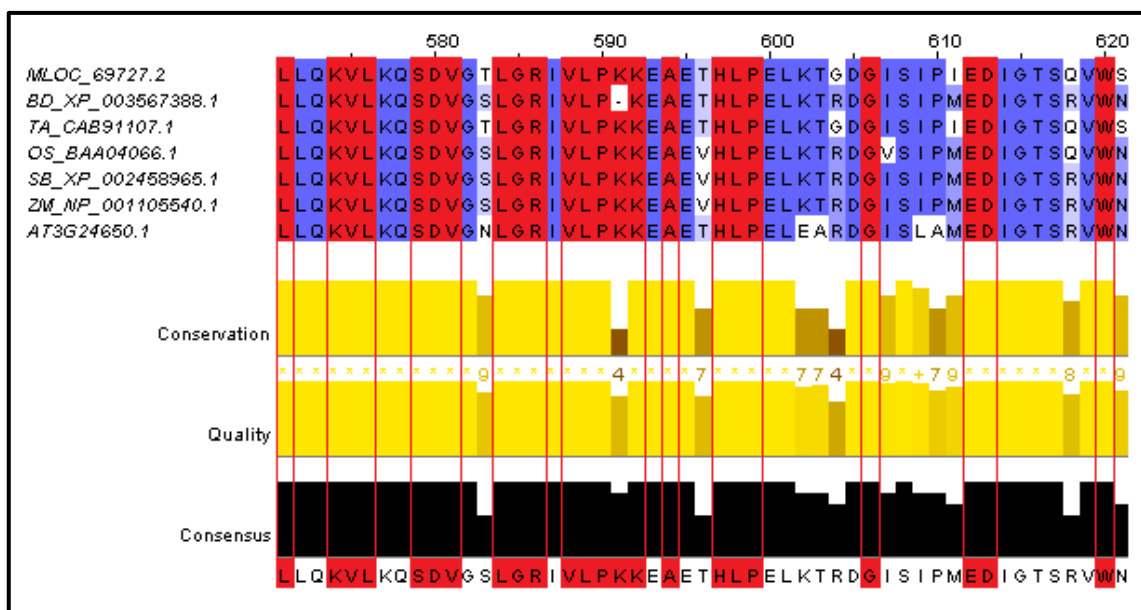
**Figure 14 - Conserved Domain Database View**

Graphical summary displays all hits detected for the input sequence. The input sequence was the RNA-Seq-Contig (contig\_137422) and it contains a B3-DNA domain, which is a specific hit. Beneath there are all detected hits listed up with description. The pairwise alignment is the conserved B3 domain and a sequence part of the RNA-Seq-Contig illustrating the similar region. The red marked amino acids are the evolutionary conserved ones within the domain.

Resulting from the BLASTx search the link to the CDD was followed directly. This is beneficial for rapidly find out about functional domains encoded by the sequence. For the concrete RNA-Seq-Contig the domain search provides one specific hit, which is described as a plant-specific B3-DNA binding domain (see Figure 14). This result is especially important because, as mentioned above, the origin candidate gene ABI3 also has a B3-DNA binding domain which is crucial for the transcriptional activity. Therefore the putative homolog in barley may also have a transcriptional function and may also be involved in barley pollen embryogenesis one way or another.

In the Figure 14 the upper part shows a graphical summary of the input sequence and the position of the detected domain. Beneath this topology all detected hits are listed. The description of the specific hit is expanded for a detailed view. Noticeably there is a consensus sequence with a pairwise alignment against a part of the input RNA-Seq-Contig. The red marked amino acids in the alignment, according to the database-specific illustration, are the functionally conserved ones.

Parallel to this the BLASTx results were observed for the previous declared related organisms, all suitable protein sequences were extracted and collected in a FASTA file. In addition to the protein sequences obtained from the NCBI a protein sequence of barley was added. This protein sequence is a previous predicted protein which was found in a set of high confidence genes from the IPK Barley Blast Server.



**Figure 15 - Conserved Region in the Multiple Protein Sequence Alignment**

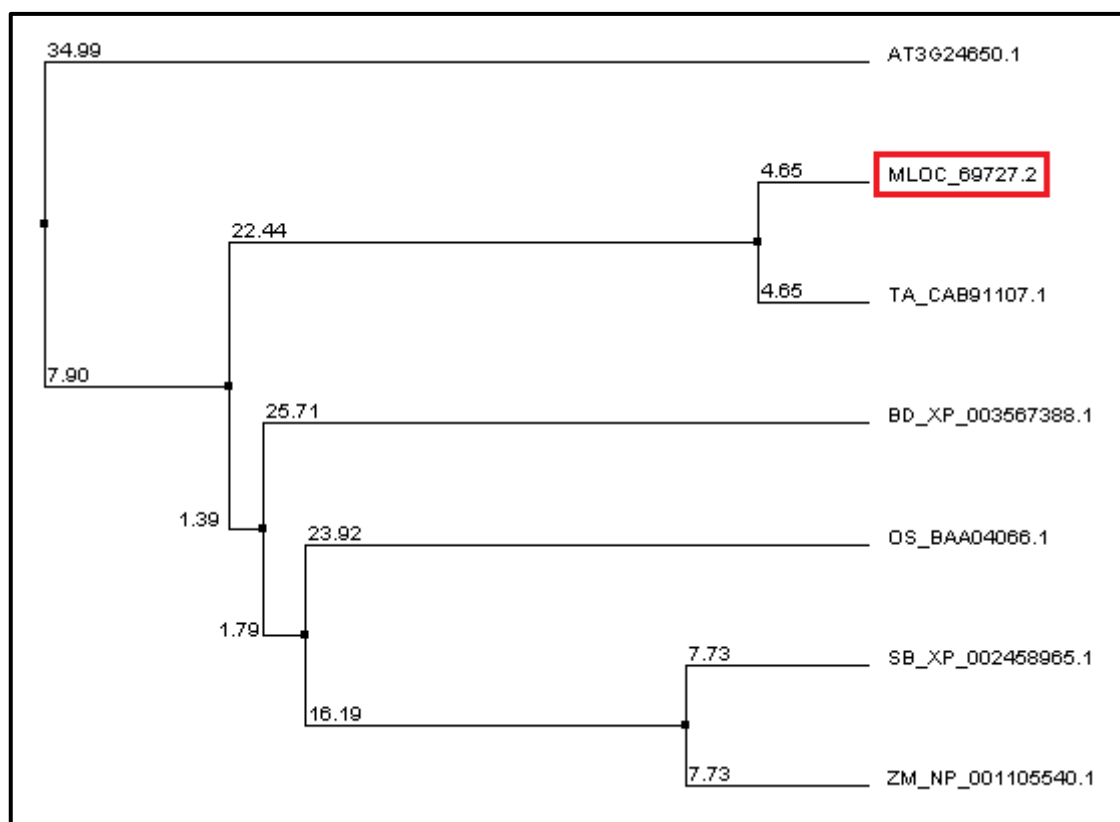
Extract from the visualization of the protein MSA in JalView. The MSA was calculated previously with ClustalW2. The sequences are amino acid sequences from related organisms to barley. The most sequences come from the NCBI. The most sequence descriptions consist of an abbreviation for the organism of origin and the accession number assigned by the NCBI. The MLOC sequence is the predicted barley protein. The sequence beginning with AT is sourced from the TAIR. MLOC: *Hordeum vulgare*, BD: *Brachypodium distachyon*, TA: *Triticum aestivum*, OS: *Oryza sativa*, SB: *Sorghum bicolor*, ZM: *Zea mays*, AT: *Arabidopsis thaliana*.

The FASTA file, containing all protein sequences available for the reference sequence, was loaded into the application ClustalW2 which produced a MSA based on the protein sequences. Subsequently the MSA was visualized in JalView. Such a protein MSA should help to find compact local regions with preferably high similarity.

The flanking regions of the MSA vary between the protein sequences. But there was a region in the middle of the MSA (see Figure 15) which was congruent for all protein sequences, except for some single amino acids. Within this matching region the conserved patterns from the B3-DNA domain detected by the CDD can be identified. This confirms the statement that this region is evolutionary conserved and due to this the B3 domain may have a defined function represented in all plant species that had been analyzed.

Furthermore the protein MSA is an excellent basis for a phylogenetic analysis, because the similarities and differences on amino acid level indicate relationships between the species. Phylogenetic relations can be easily observed with a distance tree which is calculated by JalView directly and can be seen in Figure 16. The distance tree consists of all protein sequence descriptions used in the MSA. The origin sequence is the protein predicted for barley and is marked with a red rectangle in the Figure 16. The values at the end of the branches are the distances that were calculated based on the protein MSA.

Afterwards the analysis may go back to the read coverage graph visualized in Tablet (see Figure 6). With the zoom-in option and the scroll-bars it is possible to look for the transcription start (start codon) and the reading frame based on the predicted protein sequence obtained from the BLASTx search run. In Figure 17 the start codon can be seen in the rightmost at the position 11.840 on the Morex WGS-Contig. The reading frame which is analog with the predicted protein is the second-reverse frame and the protein can be reproduced from right to left.

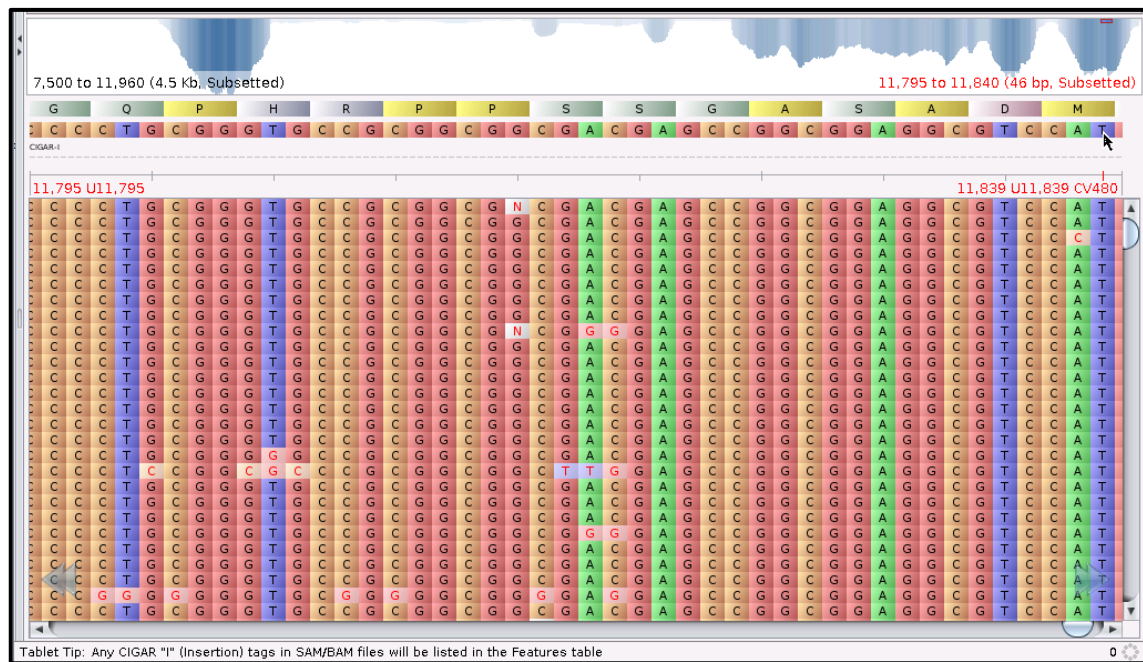


**Figure 16 - Distance Tree according to the Protein MSA**

Visualization of the distance tree in JalView. The MSA the tree is based on was calculated with ClustalW2. The values at the ends of the branches are the calculated distances. The sequences are amino acid sequences from related organisms to barley. Most sequences were extracted from the NCBI. Most sequence descriptions consists of an abbreviation for the organism of origin and the accession number assigned by the NCBI. The MLOC sequence is the predicted barley protein. The sequence beginning with AT is sourced from the TAIR. MLOC: *Hordeum vulgare*, BD: *Brachypodium distachyon*, TA: *Triticum aestivum*, OS: *Oryza sativa*, SB: *Sorghum bicolor*, ZM: *Zea mays*, AT: *Arabidopsis thaliana*.

Subsequently the stop codon can be identified by manually scrolling along the sequence in the direction detected with the previous predicted protein. This is interesting for finding the real boundaries of the gene and to compare the protein sequence with the predicted one. If the boundaries were defined the read coverage across the gene can be observed in a subset independently to the read coverage across the whole genomic sequence.

In the case of the present analyzed Morex WGS-Contig the correct start was identified but the protein sequence was only one-third congruent to the predicted sequence. Therefore the gene boundaries differ to the originally accepted start and stop codons.

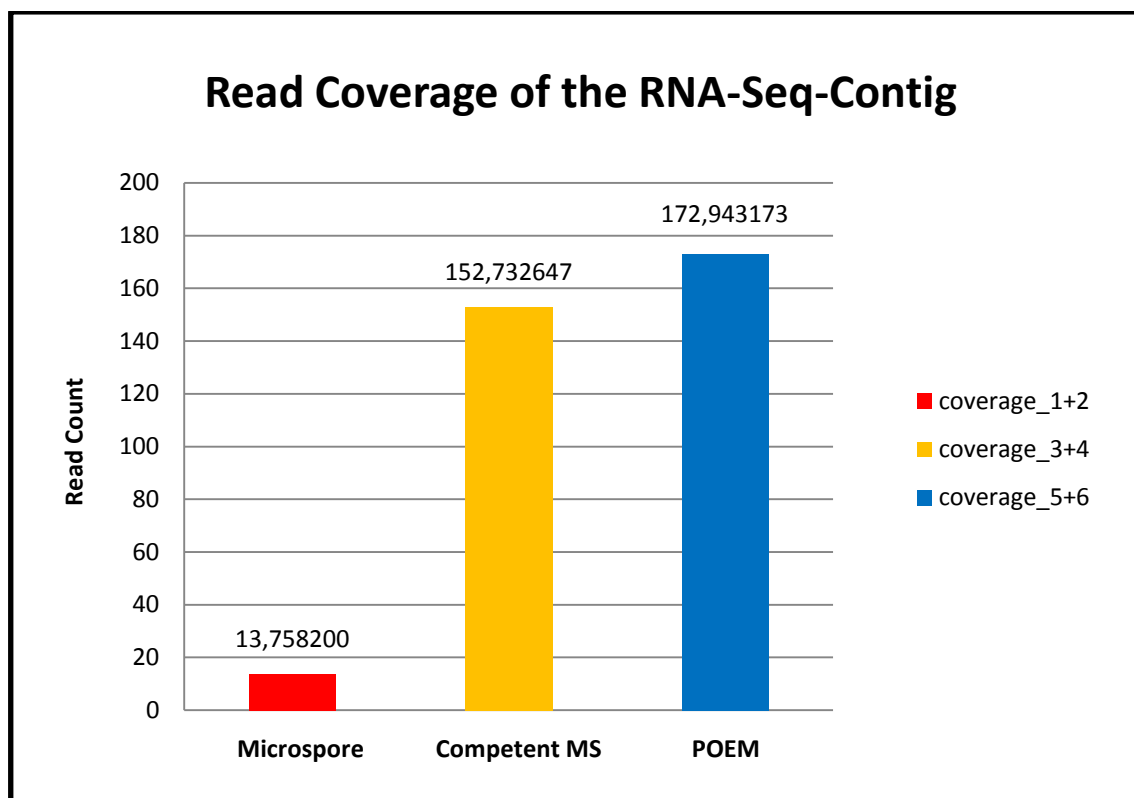


**Figure 17 - Startcodon Detection with Tablet**

Coverage graph of the Morex WGS-Contig with maximum zoom in. From right to left the protein sequence can be seen right beneath the overview window. The reading frame is the second reverse and starts with the amino acid methionine or the codon ATG at the position 11.840 on the Morex WGS-Contig. In this case the reading direction goes from right to left.

Finally the previous detailed RNA-Seq-Contig can be evaluated statistically with the read coverage. Therefore the read count values for the RNA-Seq-Contig were visualized in Microsoft Excel for the three different stages of microspore development. The read count represents no normalized values.

As it can be seen in Figure 18 the coverage of the competent microspore (3+4) is approximately ten fold higher than the coverage of the microspore in stage one (1+2). In comparison to the second stage (3+4) the coverage of the RNA-Seq-Contig in the third stage (5+6) is again slightly increased. Due to this values the ABI3 pendant in barley seem to be an interesting candidate for further molecular biological investigations.



**Figure 18 - Read Coverage of the RNA-Seq-Contig**

Averaged Number of reads re-mapped to the RNA-Seq-Contig respectively to the three stages of microspore development. The first stage (1+2) is represented by pre-mitotic microspores, which are still in regular pollen development. Second stage (3+4) is formed by embryogenesis competent microspores, which already have been stressed with the inductive treatment. And the third stage (5+6) is constituted as embryogenic pollen, one day after the treatment.

### 4.3 Summarization

After the detailed presentation of the results of one example all results should be presented in a condensed way. Therefore a table was established to give a review about the performed methods and tools as well as the decision whether they could produce meaningful results or not (see Table 2). The first column contains the gene symbols or rather abbreviations. In the second column the associated Gene-Model-IDs or accession numbers are listed, by which the genes can be identified unambiguously. The remaining columns consists of the methods and tools that have been used for the *in silico* analysis while this Bachelor thesis. The lines are filled with 'YES' if the tool achieved the planned significant result. If the results were not relevant or were not able to be achieved by the tools the cell was left empty.



**Table 2 - Summarization of the Achieved Results**

All results that could have been achieved by the tools for the twenty analyzed genes are listed up for a general view. First column contains the symbol or abbreviation of the candidate gene. Second column includes the accession numbers of the genes by which the gene can be identified unambiguously. The remaining eight columns consists of the methods and tools used while the in silico analysis and the decision whether they could produce meaningful results (YES) or not.

<b>Symbol</b>	<b>Gene Model</b>	<b>Alignment Similarity</b>	<b>Mulan</b>	<b>MultiTF</b>	<b>TriAnnot</b>	<b>Tablet</b>	<b>BLASTx</b>	<b>CDD</b>	<b>Read Count</b>
ABI3	AT3G24650.1	YES	YES	YES	YES	YES	YES	YES	YES
AGL38	AT1G65300.1								
RH9	AT3G22310.1	YES	YES		YES	YES	YES	YES	YES
BBM	AT5G17430.1	YES	YES		YES	YES	YES	YES	YES
BLH2 / SAW1	AT4G36870.1	YES	YES	YES	YES		YES		YES
BLH4 / SAW2	AT2G23760.1	YES	YES	YES	YES		YES		YES
BP / KNAT1	AT4G08150.1	YES	YES	YES	YES	YES	YES	YES	YES
EMK / AIL5	AT5G57390.1	YES	YES		YES	YES	YES	YES	YES
LEC1	AT1G21970.1	YES	YES	YES	YES	YES	YES	YES	YES
MYB115	AT5G40360.1	YES	YES	YES	YES	YES	YES	YES	YES
MYB118	AT3G27785.1	YES	YES		YES	YES	YES	YES	YES
OsH15	OS_P46609.2	YES	YES	YES	YES	YES	YES	YES	YES
PIN2	AT5G57090.1	YES	YES		YES	YES	YES	YES	YES
RKD1 (Ta)	TA_AEB26835.1	YES	YES	YES	YES	YES	YES	YES	YES
RKD4 / GRD	AT5G53040.1	YES	YES		YES	YES	YES	YES	YES
STM	AT1G62360.1	YES	YES		YES	YES	YES	YES	YES
WOX2	AT5G59340.1	YES	YES		YES	YES	YES		YES
WOX8	AT5G45980.1	YES	YES		YES	YES	YES		YES
WOX9	AT2G33880.1	YES	YES	YES	YES	YES	YES		YES
WUS	AT2G17950.1	YES	YES				YES		YES

As it can be read off Table 2 the MultiTF application was not able to detect TFBS for every set of sequences. In addition, there are not always available conserved domains that are found by the CDD. If there is no result in Tablet, as in the case of BLH2, BLH4 and WUS, there are no reads mapped onto the Morex WGS-Contig or the coverage of the RNA-Seq-Reads is too low. The rest of the methods provided adequate results that were comparable and have a force of expression. The screen shots and figures of the results can be regarded with the CD enclosed.

By comparing the Morex WGS-Contig and RNA-Seq-Contig sequences of the twenty analyzed candidate genes it attaches attention that some genes have corresponding sequences from barley in common. In Table 3 especially the genes are listed which share the genomic sequence, the transcript sequence or both.

**Table 3 - Conjugated Genes with Same Sequences**

Among the twenty candidate genes that were analyzed there are some gene with the same corresponding Morex WGS-Contigs and associated RNA-Seq-Contigs. The relation is displayed in this table through the background colors. Having the same genomic sequence did not imply having the same locus. But having the same transcript sequence can lead to the assumption that the gene is related somehow, eventually as splice alternatives.

Symbol	Gene Model	Morex WGS-Contig	RNA-Seq-Contig	CDD
BBM	AT5G17430.1	morex_contig_44345	contig_31863	AP2 - Domain
EMK / AIL5	AT5G57390.1	morex_contig_44345	contig_31863	AP2 - Domain
BLH2 / SAW1	AT4G36870.1	morex_contig_136249	contig_137432	
BLH4 / SAW2	AT2G23760.1	morex_contig_136249	contig_137432	
BP / KNAT1	AT4G08150.1	morex_contig_1561605	contig_34186	KNOX1 / KNOX2 / HOX
OsH15	OS_P46609.2	morex_contig_1561605	contig_34186	KNOX1 / KNOX2 / HOX
STM	AT1G62360.1	morex_contig_1561605	contig_34186	KNOX1 / KNOX2 / HOX
MYB115	AT5G40360.1	morex_contig_42106	contig_109134	SANT - Domain
MYB118	AT3G27785.1	morex_contig_42106	contig_109134	SANT - Domain
TaRKD1	TA_AEB26835.1	morex_contig_65307	contig_76112	RWP-RK - Domain
RKD4 / GRD	AT5G53040.1	morex_contig_495	contig_76112	
WOX2	AT5G59340.1	morex_contig_47071	contig_85409	
WUS	AT2G17950.1	morex_contig_1585113	contig_85409	

If two genes are located on the same genomic contig this must not imply that they have the same locus. But if they also have the same transcript sequence the assumption that the genes are related somehow can be made. Maybe due to this, the genes are splice alternatives.

The last two examples in Table 3 are especially interesting because they are not located on the same genomic region but have the same transcript sequence. This can be explained with an assembling error due to a confusable similarity of two genomic regions. The assembling of the RNA-Seq-Contigs was produced without a reference

sequence and therefore the RNA-Seq-Reads collapsed into only one contig because of the sequence similarity. These genes (TaRKD1 and RKD4 as well as WOX2 and WUS) can be interpreted as two different pendants in barley in each case. The genes which share both the genomic and the transcript sequence may be just one pendant in barley originated from two genes through evolutionary development. Otherwise one specific genomic locus can encode one transcript sequence which is spliced afterwards and gives rise to two different proteins in barley.

## **5 Discussion**

These twenty genes, analyzed while the present Bachelor thesis, are an adequate number for the task settings. But some of these twenty genes have the same associated genomic and transcript sequences in barley (see Table 3). Therefore it cannot be concluded unambiguously which gene is represented with an appropriate putative homolog in barley. For discrimination, between eventually different structurally composed or closely spaced genes, further analysis will be necessary.

To summarize the results in an appropriate manner a table was established, displaying which tool provides significant results, associated to the twenty analyzed candidate genes (see Table 2). For every gene, except for the AGL38 gene, an alignment similarity was ascertained and visualized with Mulan. Additionally a bar chart was performed for all the nineteen genes, illustrating the read count across the three stages of microspore development. The BLASTx search determined similar protein sequences in previously selected related organisms, again for all nineteen candidate genes. Unfortunately the remaining tools were not able to produce significant results for every gene.

The protein sequences for the MSAs were mostly predicted sequences and were used from public databases. That means the sequences are not validated or revised. If the predicted sequence matches with the other sequences, the prediction seems to be reliable. But the other way around, there are more problems with the interpretation if the protein MSA did not show analogies. For example the protein sequence, associated to the RKD1 gene from *Triticum aestivum*, was predicted as a high confidence gene in barley. But the MSA shows just a short region with concordance between all input sequences. This may be caused by inaccurate protein prediction and is additionally complicated through different protein lengths. In contrast the protein sequences, which were detected corresponding to the ABI3 gene from *Arabidopsis thaliana*, helped to find the conserved region of the B3 domain, which is available in all analyzed, related organisms (see Figure 15).

The coverage graphs of the RNA-Seq-Reads against the Morex WGS-Contigs were performed in Tablet with all reads from the three stages of microspore development. Unfortunately through this combination of reads the real expression levels of the genes cannot be observed. With a continuation of the analysis a new mapping will be worthy, by which the RNA-Seq-Reads are mapped against the WGS assembly, separately dedicated to the three stages of microspores. This may illustrate differences between the expression levels of the three stages and additionally may point out splice alternatives.

With regard to the already mentioned problem of predicted protein sequences Tablet is able to prove or rebut the prediction and the quality of the protein sequences. The Morex WGS-Contig, associated to the ABI3 gene from *Arabidopsis thaliana*, show that the second reverse reading frame is the used direction of transcription (see Figure 17). Comparing the protein sequence predicted from Tablet with the high confidence gene from barley, only two third of the sequences are completely homolog. At a certain position in Tablet the protein sequence differs completely from the predicted sequence. Furthermore the stop codon is following right after the inhomogeneity. The irregularity can only be identified through manual examination and may be caused by different prediction algorithms or through varying WGS-Contig sequences.

The last issue with Tablet is the problem of false positive mapped reads. As mentioned in section 3.8 there are reads with unmapped mates in the middle of the contig (see Figure 8). This is usually impossible because of the insert size of 200 bp due to the Illumina run. These false positive mapped reads manipulate the coverage and therefore pretend a higher read coverage. But these red marked reads (see Figure 8) with an unmapped mate indicate that reads, which do not belong to the transcribed region, can map anyway. This is because the mapping algorithms rely on position-specific sequence similarity instead of implying the topology.

In the end the read count of the RNA-Seq-Contigs in comparison of the three stages of microspore development were visualized in a bar chart with Excel for a statistical

evaluation. These read count values are the non-normalized coverage of the transcripts. The normalization was not performed, because the number of reads was comparable for the three stages. The problem is that different splice alternatives were not observed and varying exon lengths were ignored. In further analysis the read count values should be normalized into FPKM metric to include all aberrations and variances.

## **6 Summary and Outlook**

### **6.1 Summary**

The main purpose of this Bachelor thesis was to find and compile comprehensive information about appropriate putative homologs in barley to previous given genes due to an educated guess from the conductor. The genes identified in barley should serve as trigger-genes to increase the switch of microspores to pollen embryogenesis in the very end.

First the homologous genes to the previously known were searched and identified in barley. Therefore the basic BLAST algorithm was used. Second step was to analyze the detected genomic and associated transcript sequences *in silico*, to provide a suitable structural and functional annotation.

The annotation was performed with several bioinformatics tools and their individual results were brought together to achieve a general view. Third the results of one representative example, chosen from the total of analyzed genes, was presented as detailed as possible. And at last all results achieved by all *in silico* analysis were clustered in tables for a convenient overview.

For nineteen genes out of the twenty previous selected the annotations could be made and provided mostly significant results. All in all for every gene a glance could be achieved about what exonic structure they have and where the conserved or repetitive regions can be located.

But these annotations just provide an overview about the structure and function. For an exact comprehension of the gene with its interactions, pathways and movements there are more detailed analyses required.

In the end the results provide a suitable basis for further *in silico* or *in vitro* analysis. The results contribute to a significantly improved understanding of barley pollen embryogenesis.

## 6.2 Outlook

Due to the discussion there are some open issues for further analysis. First it would be interesting to make a new mapping of the RNA-Seq-Reads against the Morex WGS assembly, separately dedicated to the three stages of microspore development. Such a mapping could be performed for example with the BWA-SW algorithm which is a new implemented, large-scale, rapid and accurate aligner [Li, H.; Durbin, R. (2010)].

Besides a normalization of the coverage values may be motivating. The software Cufflinks could be an appropriate application because it is able to normalize RNA-Seq samples and additionally provide information about the differential expression and regulation of the RNA-Seq reads [Trapnell, C. et al. (2012)].

Both would help to identify structural differences and expression level changes between the different microspore stages. And both are bioinformatical methods that require an improved knowledgebase for the tools and algorithms. Predictions about the expression levels and structure dissimilarities may answer the questions about the genes which have putatively the same barley homologs.

Additional the second approach, displayed in the pipeline in section 3.2 (see Figure 4), may be followed in further analysis. This alternative approach may detect entirely new genes involved in the initiation of pollen embryogenesis in barley. Therefore the RNA-Seq-Contigs can be clustered due to their expression pattern, as it was partly realized in the internship report, which is associated to the present Bachelor thesis. Such a clustering could be achieved by differentiating due to the fold change [Mutch, D. M. et al. (2002)].

Another advance could be the observation of possible relations or co-operations between the unknown RNA-Seq-Contigs and the annotated genes presented in this present Bachelor thesis. These investigations could be based on already existing integrated protein-protein interaction networks. For example the tool ANAP



(Arabidopsis Network Analysis Pipeline) illustrates interactions between gene products from *Arabidopsis thaliana* in an interactive and intuitive viewer [Wang, C. et al. (2012); URL-21]. These can serve as a basic concept for further investigations with the putatively homologous genes in barley. For example the interaction between the ABI3 gene from *Arabidopsis thaliana* with the LEC1 gene was mentioned previously. This interaction, in addition with further network entries, could be visualized with the ANAP (see Appendix 3) and may be this is assignable to the analyzed putative homologs in barley. This would be a huge amount of manual comparison and analysis but would help to understand the protein interaction and involvements.

In the very end it is of course the most interesting part to find out if one of these genes in barley is able to increase the number of microspores switching to pollen embryogenesis. These investigations can be performed through molecular biological methods. Such analyses of these specifically selected genes may help to improve the knowledge and understanding of barley pollen embryogenesis and the scientific application.

**List of References**

**Garber, M. et al. (2011):** Computational methods for transcriptome annotation and quantification using RNA-seq. *Nature Methods*, 8: 469-477

**Jeong, S. et al. (2011):** The RWP-RK Factor GROUNDED Promotes Embryonic Polarity by Facilitating YODA MAP Kinase Signaling. *Current Biology*, 21: 1268-1276

**Korf, I. et al. (2003):** BLAST. *O'Reilly&Associates, Inc.*

**Koszegi, D. et al. (2011):** Members of the RKD Transcription Factor Family Induce an Egg Cell-Like Gene Expression Program. *The Plant Journal*, 67: 280-291

**Lee, J.-H. et al. (2008):** Early Sexual Origins of Homeoprotein Heterodimerization and Evolution of the Plant KNOX/BELL Family. *Cell*, 133:829-840

**Leroy, P. et al. (2012):** TriAnnot: a versatile and high performance pipeline for the automated annotation of plant genomes. *Frontiers in Plant Science*, 3: 1-14 DOI 10.3389/fpls.2012.00005

**Li, H.; Durbin, R. (2010):** Fast and Accurate Long-Read Alignment with Burrows-Wheeler transform. *Bioinformatics*, 26: 589-595

**Lippmann, R. (2012):** Physiologische und bioanalytische Untersuchungen während der Pollenembryogenese von *Hordeum vulgare* [Dissertation] Halle: Fakultät Biowissenschaften, Martin-Luther-Universität Halle-Wittenberg

**Magnani, E.; Hake, S. (2008):** KNOX Lost the OX: The Arabidopsis KNATM Gene defined a Novel Class of KNOX Transcriptional Regulators Missing the Homeodomain. *The Plant Cell*, 20: 875-887

- Maraschin, S. F. et al. (2006):** cDNA array analysis of stress-induced gene expression in barley androgenesis. *Physiologia Plantarum*, 127: 535-550
- Marchler-Bauer, A. et al. (2009):** CDD: specific functional annotation with the Conserved Domain Database. *Nucleic Acid Research*, 37: D205-D210
- Milne, I. et al. (2010):** Tablet - next generation sequence assembly visualization. *Bioinformatics*, 26:401-402
- Milne, I. et al. (2012):** Using Tablet for visual exploration of second-generation sequencing data. *Briefings in Bioinformatics*, 14: 193-202
- Mutch, D. M. et al. (2002):** The limit fold change model: A practical approach for selecting differentially expressed genes from microarray data. *BMC Bioinformatics*, 3:7
- Nagalakshmi, U. et al. (2010):** RNA-Seq: A Method for Comprehensive Transcriptome Analysis. *Current Protocols in Molecular Biology*, 89: 4.11.1-4.11.13
- Oestereich, B. (1999):** Objektorientierte Softwareentwicklung, Analyse und Design mit der Unified Modeling Language. 4.Auflage R. Oldenburg Verlag, München
- Ovcharenko, I. et al. (2005):** Mulan: Multiple-sequence local alignment and visualization for studying function and evolution. *Genome Research*, 15: 184-194
- Reynolds, T. L. (1997):** Pollen embryogenesis. *Plant Molecular Biology*, 33: 1-10
- Sentoku, N. et al. (1999):** Regional Expression of the Rice KN1-Type Homeobox Gene Family during Embryo, Shoot and Flower Development. *The Plant Cell*, 11: 1651-1663

**Silva, T. D. (2012):** Microspore Embryogenesis. *Embryogenesis*, Dr. Ken-Ichi Sato (Ed.), ISBN: 978-953-51-0466-7, *InTech*, Available from:

<http://www.intechopen.com/books/embryogenesis/microspore-embryogenesis>

**Tamaoki, M. et al. (1995):** Alternative RNA products from a rice homeobox gene. *The Plant Journal*, 7: 927-938

**The International Barley Genome Sequencing Consortium (2012):** A physical, genetic and functional sequence assembly of the barley genome. *Nature*, 491: 711-717

**Trapnell, C. et al. (2012):** Differential gene and transcript expression analysis of RNA-Seq experiments with TopHat and Cufflinks. *Nature Protocols*, 7:562-578

**Tsuwamoto, R. et al. (2010):** Arabidopsis EMBRYOMAKER encoding an AP2 domain transcription factor plays a key role in developmental change from vegetative to embryonic phase. *Plant Molecular Biology*, 73: 481-492

**URL-1** (Access: 10.06.2013): Unknown Author, Barley (*Hordeum vulgare* L.). Available from: [http://plantgenetics.lu.lv/html/miezi\\_en.html](http://plantgenetics.lu.lv/html/miezi_en.html)

**URL-2** (Access: 14.06.2013): Unknown Author, illumina, Genome Analyzer Iix. Available from: [http://www.illumina.com/systems/genome\\_analyzer\\_iix.ilmn](http://www.illumina.com/systems/genome_analyzer_iix.ilmn)

**URL-3** (Access: 04.04.2013): BLAST® Help [Internet]. Bethesda (MD): National Center for Biotechnology Information (US); 2008-. Available from: <http://www.ncbi.nlm.nih.gov/books/NBK1762/>

**URL-4** (Access: 24.06.2012): IPK Barley BLAST Server. Basic Search. Available from: <http://webblast.ipk-gatersleben.de/barley/>

- URL-5** (Access: 25.06.2012): Unknown Author, Translated BLAST: blastx. Available from:  
[http://blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastx&BLAST\\_PROGRAMS=blastx&PAGE\\_TYPE=BlastSearch&SHOW\\_DEFAULTS=on&LINK\\_LOC=blasthome](http://blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastx&BLAST_PROGRAMS=blastx&PAGE_TYPE=BlastSearch&SHOW_DEFAULTS=on&LINK_LOC=blasthome)
- URL-6** (Access: 05.04.2013): Unknown Author, ClustalW2, Help & Documentation. Available from: <http://www.ebi.ac.uk/Tools/msa/clustalw2/help/>
- URL-7** (Access: 05.04.2013): Unknown Author, 2Can Support Portal, Clustalw2 multiple sequence alignment. Available from:  
<http://www.ebi.ac.uk/2can/tutorials/protein/clustalw.html>
- URL-8** (Access: 19.04.2013): Unknown Author, Conserved Domains and Protein Classification. Available from:  
[http://www.ncbi.nlm.nih.gov/Structure/cdd/cdd\\_help.shtml](http://www.ncbi.nlm.nih.gov/Structure/cdd/cdd_help.shtml)
- URL-9** (Access: 02.07.2013): Unknown Author, Tablet - Next Generation Sequence Assembly Visualization. Available from: <http://bioinf.scri.ac.uk/tablet/index.shtml>
- URL-10** (Access: 02.07.2013): Unknown Author, mRNA Sequencing, Illumina RNA-Seq Process. Available from: <http://www.oceanridgebio.com/mrna-sequencing.html>
- URL-11** (Access: 02.07.2013): Carolyn Elya, Next-gen sequencing. Available from:  
[http://www.eisenlab.org/FunFly/?page\\_id=24](http://www.eisenlab.org/FunFly/?page_id=24)
- URL-12** (Access: 03.07.2013): Unknown Author, Nahrung, Energie und Rohstoffe aus Pflanzen. Portraits der Gerste. Available from:  
<http://www.bundesregierung.de/Content/DE/Artikel/WissenschaftWohlstand/2008-04-01-hightech-serie-pflanzen-hauptartikel.html>

**URL-13** (Access: 03.07.2013): Unknown Author, Geschichte. Seit Jahrtausenden auf unserem Speiseplan. Available from: <http://www.dieckmann-cereals.de/index.php/betagerste-von-a-z/geschichte/>

**URL-14** (Access: 03.07.2013): Thomas Reich, Scubavision. Sommer. Available from: <http://www.scubavision.de/foto-gerste-1275-473.html>

**URL-15** (Access: 03.07.2013): M.J. Edney, The road to quality hulless malting barley - Where to now? History of hulless barley for malting in Canada. Available from: <http://www.grainscanada.gc.ca/research-recherche/edney/barley-orge/hb-ogn-eng.htm>

**URL-16** (Access: 09.07.2013): Unknown Author, NCBI, DNA & RNA, Sequence Read Archive (SRA). Available from: <http://www.ncbi.nlm.nih.gov/sra>

**URL-17** (Access: 09.07.2012): H. Quesneville, Triannot Pipeline, Triannot - v3.8. Available from: <http://wheat-urgi.versailles.inra.fr/Tools/Triannot-Pipeline>

**URL-18** (Access: 17.07.2013): NCBI DCODE.org Comparative Genomics Developments, Mulan. Available from: <http://mulan.dcode.org/>

**URL-19** (Access: 17.07.2013): BIOBASE Biological Databases, TRANSFAC® Transcription Factor Binding Sites. Available from: <http://www.biobase-international.com/product/transcription-factor-binding-sites>

**URL-20** (Access: 21.06.2013): The Arabidopsis Information Resource - TAIR Gene Search Results. Available from: [http://www.arabidopsis.org/servlets/Search?action=new\\_search&type=gene](http://www.arabidopsis.org/servlets/Search?action=new_search&type=gene)

**URL-21** (Access: 14.08.2013): ANAP - Arabidopsis Network Analysis Pipeline. Available from: [http://gmdd.shgmo.org/Computational-Biology/ANAP/ANAP\\_V1.1/](http://gmdd.shgmo.org/Computational-Biology/ANAP/ANAP_V1.1/)

**Vollbrecht, E. et al. (1991):** The Developmental Gene Knotted-1 is a Member of a Maize Homeobox Gene Family. *Nature*, 350: 241-243

**Waki, T. et al. (2011):** The Arabidopsis RWP-RK Protein RKD4 Triggers Gene Expression and Pattern Formation in Early Embryogenesis. *Current Biology*, 21:1277-1281

**Wang, C. et al. (2012):** ANAP: An Integrated Knowledge Base for Arabidopsis Protein Interaction Network Analysis. *Plant Physiology*, 158:1523-1533

**Wang, Z. et al. (2009):** RNA-Seq: A Revolutionary Tool for Transcriptomics. *Nature*, 10: 57-63

## Appendix

### Appendix 1

Symbol	Full Name	Gene Model ID	Available Description	Literature
<b>Embryo identity</b>				
ABI3	ABSCISIC ACID INSENSITIVE 3	AT3G24650.1	[Arabidopsis thaliana] <a href="#">Homologous to the maize transcription factor Viviparous-1</a> . Full length ABI3 protein binds to the highly conserved RY motif [DNA motif CATGCA(TG)], present in many seed-specific promoters, and the B3 domains of this transcription factor is necessary for the specific interaction with the RY element. Transcriptional activity of ABI3 requires the B3 DNA-binding domain and an activation domain. In addition to the known N-terminal-located activation domain, a second transcription activation domain was found in the B1 region of ABI3. ABI3 is essential for seed maturation. <a href="#">Regulator of the transition between embryo maturation and early seedling development</a> . Putative seed-specific transcriptional activator. ABI3 is a central regulator in ABA signaling and is unstable in vivo. It interacts with AIP2 and can be polyubiquitinated by AIP2 in vivo. Based on double mutant analyses, <a href="#">ABI3 interacts genetically with both FUS3 and LEC1</a> and is involved in controlling accumulation of chlorophyll and anthocyanins, sensitivity to abscisic acid, and expression of the members of the 12S storage protein gene family. In addition, both FUS3 and LEC1 regulate positively the abundance of the ABI3 protein in the seed. Alternative splicing of ABI3 is developmentally regulated by SUA (AT3G54230). [URL-20]	



Symbol	Full Name	Gene Model ID	Available Description	Literature
<b>AGL15</b>	AGAMOUS-LIKE 15	AT5G13790.1	[Arabidopsis thaliana] AGL15 (AGAMOUS-Like 15) is a <a href="#">member of the MADS domain family of regulatory factors</a> . Although AGL15 is <a href="#">preferentially expressed during embryogenesis</a> , AGL15 is also expressed in leaf primordia, shoot apical meristems and young floral buds, suggesting that AGL15 may play a role during post-germinative development. Transgenic plants that ectopically express AGL15 show delays in the transition to flowering, perianth abscission and senescence and fruit and seed maturation. Role in embryogenesis and gibberellic acid catabolism. Targets B3 domain transcription factors that are key regulators of embryogenesis. [URL-20]	
<b>AGL38 / PHE2</b>	AGAMOUS-LIKE 38	AT1G65300.1	[Arabidopsis thaliana] Encodes PHERES2, a homolog of PHERES1. PHERES1 and PHERES2 are both target genes of the FIS Polycomb group complex but only PHERES1 is regulated by genomic imprinting, which is likely caused by the presence of repeat sequences in the proximity of the PHERES1 locus. [URL-20]	
<b>AP2</b>	APETALA 2	AT4G36920.1	[Arabidopsis thaliana] Encodes a floral homeotic gene, a member of the AP2/EREBP (ethylene responsive element binding protein) class of transcription factors and is involved in the specification of floral organ identity, establishment of floral meristem identity, suppression of floral meristem indeterminacy, and development of the ovule and seed coat. AP2 also has a role in controlling seed mass. Dominant negative allele I28, revealed a function in meristem maintenance-mutant meristems are smaller than normal siblings. AP2 appears to act on the WUS-CLV pathway in an AG independent manner. [URL-20]	
<b>BBM</b>	BABY BOOM	AT5G17430.1	[Arabidopsis thaliana] Encodes an AP2-domain containing protein similar to ANT. <a href="#">Expressed in embryos</a> and lateral root primordium. [URL-20]	

Symbol	Full Name	Gene Model ID	Available Description	Literature
EMK1 / AIL5	EMBRYO- MAKER / AINTEGUMENT A-LIKE 5	AT5G57390.1	[Arabidopsis thaliana] AP2-family Embryomaker. <a href="#">BBM and AIL5 homologue. Switches vegetative to embryonic development.</a> Arabidopsis EMBRYOMAKER encoding an AP2 domain transcription factor plays a key role in developmental change from vegetative to embryonic phase. Encodes a <a href="#">member of the AP2 family of transcriptional regulators. May be involved in germination and seedling growth.</a> Mutants are resistant to ABA analogs and are resistant to high nitrogen concentrations -essential for the developmental transition between the embryonic and vegetative phases in plants. Overexpression results in the formation of somatic embryos on cotyledons. It is also required to maintain high levels of PIN1 expression at the periphery of the meristem and modulate local auxin production in the central region of the SAM which underlies phyllotactic transitions. [URL-20]	Tsuwamoto, R.; Yokoi, S.; Takahata, T. (2010)
LEC1	LEAFY COTYLEDON 1	AT1G21970.1	[Arabidopsis thaliana] <a href="#">Transcriptional activator of genes required for both embryo maturation and cellular differentiation.</a> Sequence is similar to HAP3 subunit of the CCAAT-box binding factor. HAP3 subunit is divided into three domains: an amino-terminal A domain, a central B domain, and a carboxyl-terminal C domain. LEC1 shared high similarity with other HAP3 homologs only in central, B domain. LEC1 is required for the specification of cotyledon identity and the completion of embryo maturation. It was sufficient to induce embryogenic programs in vegetative cells, suggesting that LEC1 is a major embryonic regulator that mediates the switch between embryo and vegetative development. Mutants are desiccation intolerant, have trichomes on cotyledons and exhibit precocious meristem activation. Levels of the ABI3 and FUS3 transcripts were significantly reduced in developing siliques of the lec1-1 mutants, indicating that LEC1 down-regulates FUS3 and ABI3. When LEC1 is overexpressed from an inducible promoter, the expression of numerous genes involved in fatty acid biosynthesis is increased suggesting a role in positive regulation of FA biosynthesis. [URL-20]	

Symbol	Full Name	Gene Model ID	Available Description	Literature
MYB115	MYB DOMAIN PROTEIN 115	AT5G40360.1	[Arabidopsis thaliana] putative transcription factor [URL-20]	
MYB118	MYB DOMAIN PROTEIN 118	AT3G27785.1	[Arabidopsis thaliana] putative transcription factor [URL-20]	
<b>Gametophyte</b>				
RKD4/ GRD	RWP-RK DOMAIN-CONTAINING 4	AT5G53040.1	[Arabidopsis thaliana] Encodes GROUNDED (GRD), a putative RWP-RK-type transcription factor broadly expressed in early development. <a href="#">GRD promotes zygote elongation and basal cell fates.</a> [URL-20]	[Waki, T. et al. (2011)] and [Jeong, S.; Palmer, T.M.; Lukowitz, W. (2011)]
RKD1	RWP-RK DOMAIN-CONTAINING PROTEIN GENE	JF714946.1	[Triticum aestivum]	Koszegi, D. et al. (2011):
<b>Cell cycle</b>				
CDC2A (CDKA1)	CYCLIN-DEPENDENT KINASE A 1	AT3G48750.1	[Arabidopsis thaliana] A-type cyclin-dependent kinase. Together with its specific inhibitor, the Kip-related protein, KRP2 they regulate the mitosis-to-endocycle transition during leaf development. Dominant negative mutations abolish cell division. Loss of function phenotype has reduced fertility with failure to transmit via pollen. Pollen development is arrested at the second mitotic division. Expression is regulated by environmental and chemical signals. Part of the promoter is responsible for expression in trichomes. Functions as a positive regulator of cell proliferation during development of the male gametophyte, embryo and endosperm. Phosphorylation of threonine 161 is required for activation of its associated kinase. [URL-20]	

Symbol	Full Name	Gene Model ID	Available Description	Literature
<b>Homeobox proteins</b>				
<b>BP/ KNAT-1</b>	BREVIPEDICELL US	AT4G08150.1	[Arabidopsis thaliana] A member of <a href="#">class I knotted1-like homeobox gene family (together with KNAT2)</a> . Similar to the knotted1 (kn1) homeobox gene of maize. Normally expressed in the peripheral and rib zone of shoot apical meristem but not in the leaf primordia. It is also expressed in the fourth floral whorl, in the region that would become style, particularly in the cell surrounding the transmitting tissue. No expression was detected in the first three floral whorls. Expression is repressed by auxin and AS1 which results in the promotion of leaf fate. [URL-20]	Lee, J.-H. et al. (2008)
<b>ChrGSM1</b>	GAMETE- SPECIFIC MINUS 1 GENE	EU029996.1	[Chlamydomonas reinhardtii]	Lee, J.-H. et al. (2008)
<b>ChrZSP2-1</b>	ZYGOTE- SPECIFIC LECTIN-LIKE PROTEIN GENE zsp2-1 allele	AF053098.1	[Chlamydomonas reinhardtii]	
<b>ChrZSP2-2</b>	ZYGOTE- SPECIFIC LECTIN-LIKE PROTEIN GENE zsp2-2 allele	AF053099.1	[Chlamydomonas reinhardtii]	
<b>KNAT3</b>	KNOTTED1-LIKE HOMEBOX GENE 3	AT5G25220.1	[Arabidopsis thaliana] A member of class II knotted1-like homeobox gene family (together with KNAT4 and KNAT5). Expressed in: hypocotyl-root boundary, anther-filament junction in flowers, ovule-funiculus and peduncle-silique boundaries, petioles and root. <a href="#">Light-regulated expression with differential response to red/far-red light</a> . KNAT3 promoter activity showed cell-type specific pattern along longitudinal root axis, restricted mainly to the differentiation zone of the root, namely in the cortex and	Lee, J.-H. et al. (2008)

Symbol	Full Name	Gene Model ID	Available Description	Literature
			pericycle. Not detected in lateral root primordia. [URL-20]	
<b>KNAT4</b>	KNOTTED1-LIKE HOMEODOMAIN GENE 4	AT5G11060.1	[Arabidopsis thaliana] A member of Class II KN1-like homeodomain transcription factors (together with KNAT3 and KNAT5), with greatest homology to the maize knox1 homeobox protein. <a href="#">Expression regulated by light</a> . Detected in all tissues examined, but most prominent in leaves and young siliques. Transient expression of GFP translational fusion protein suggests bipartite localization in nucleus and cytoplasm. KNAT4 promoter activity showed cell-type specific pattern along longitudinal root axis; GUS expression pattern started at the elongation zone, predominantly in the phloem and pericycle cells, extending to endodermis toward the base of the root. [URL-20]	
<b>KNAT5</b>	KNOTTED1-LIKE HOMEODOMAIN GENE 5	AT4G32040.1	[Arabidopsis thaliana] A member of Class II KN1-like homeodomain transcription factors (together with KNAT3 and KNAT4), with greatest homology to the maize knox1 homeobox protein. Regulates photomorphogenic responses and represses late steps in gibberellin biosynthesis. KNAT5 promoter activity showed cell-type specific pattern along longitudinal root axis, primarily in the epidermis of the distal end of primary root elongation zone. [URL-20]	
<b>KNATM</b>	KNOX ARABIDOPSIS THALIANA MEINOX	AT1G14760.1	[Arabidopsis thaliana] Encodes a novel Arabidopsis KNOX gene that encodes a MEINOX domain but lacks the homeodomain and interacts with TALE-class homeodomain proteins to modulate their activities. [URL-20]	Magnani, E.; Hake, S. (2008)
<b>OsH15</b>	OsH15 GENE KNOX CLASS 1 HOMEODOMAIN PROTEIN	AB262805.1	[Oryza sativa]	[Lee, J.-H. et al. (2008)] and [Sentoku, N. et al. (1999)]
<b>OsH45</b>	OSH45 GENE	D49704.2	[Oryza sativa]	[Lee, J.-H. et al. (2008)] and [Tamaoki, M. et al. (1995)]

<u>Symbol</u>	<u>Full Name</u>	<u>Gene Model ID</u>	<u>Available Description</u>	<u>Literature</u>
<b>PmSKN1</b>		AAD00691.1	[Picea mariana] homeobox transcription factor	Lee, J.-H. et al. (2008)
<b>STM</b>	SHOOT MERISTEMLESS	AT1G62360.1	[Arabidopsis thaliana] Class I knotted-like homeodomain protein that is required for shoot apical meristem (SAM) formation during embryogenesis and for SAM function throughout the lifetime of the plant. Functions by preventing incorporation of cells in the meristem center into differentiating organ primordia. [URL-20]	
<b>StPOTH1</b>		U65648.1	[Solanum tuberosum] homeodomain protein POTH1	Lee, J.-H. et al. (2008)
<b>WOX2</b>	WUSCHEL RELATED HOMEODOMAIN 2	AT5G59340.1	[Arabidopsis thaliana] Encodes a <a href="#">WUSCHEL-related homeobox gene family member</a> with 65 amino acids in its homeodomain. Proteins in this family contain a sequence of eight residues (TLPLFPMH) downstream of the homeodomain called the WUS box. WOX2 has a putative Zinc finger domain downstream of the homeodomain. Transcripts are expressed in the egg cell, the zygote and the apical cell lineage and are reduced in met3-1 early embryos. This gene is necessary for cell divisions that form the apical embryo domain. [URL-20]	
<b>WOX8</b>	WUSCHEL RELATED HOMEODOMAIN 8	AT5G45980.1	[Arabidopsis thaliana] Arabidopsis thaliana WOX8 protein. Contains similarity to homeodomain transcription factor. <a href="#">Positively regulates early embryonic growth</a> . Together with CLE8 it forms a signaling module that promotes seed growth and overall seed size. [URL-20]	
<b>WOX9</b>	WUSCHEL RELATED HOMEODOMAIN 9	AT2G33880.1	[Arabidopsis thaliana] <a href="#">Encodes a protein with similarity to WUS type homeodomain protein</a> . Required for meristem growth and development and acts through positive regulation of WUS. Loss of function phenotypes includes embryo lethality, hyponastic cotyledons, reduced root development and smaller meristems. Phenotypes can be rescued by addition of sucrose in the growth media. Overexpression can partially rescue the triple mutant cytokinin receptor phenotype suggesting HB-3 is a	

<u>Symbol</u>	<u>Full Name</u>	<u>Gene Model ID</u>	<u>Available Description</u>	<u>Literature</u>
			downstream effector of cytokinin signaling.	
<b>WUS</b>	WUSCHEL	AT2G17950.1	[Arabidopsis thaliana] Homeobox gene controlling the stem cell pool. Expressed in the stem cell organizing center of meristems. Required to keep the stem cells in an undifferentiated state. Regulation of WUS transcription is a central checkpoint in stem cell control. The size of the WUS expression domain controls the size of the stem cell population through WUS indirectly activating the expression of CLAVATA3 (CLV3) in the stem cells and CLV3 repressing WUS transcription through the CLV1 receptor kinase signaling pathway. Repression of WUS transcription through AGAMOUS (AG) activity controls stem cell activity in the determinate floral meristem. Binds to TAAT element core motif. WUS is also involved in cell differentiation during anther development. [URL-20]	
<b>ZmKN1</b>	KNOTTED-1 GENE	X61308.1	[Zea mays]	[Lee, J.-H. et al. (2008)] and [Vollbrecht, E. et al. (1991)]
<b>BELL-Family</b>				
<b>ATH1</b>	HOMEBOX GENE 1	AT4G32980.1	[Arabidopsis thaliana] Encodes transcription factor involved in photomorphogenesis. Regulates gibberellin biosynthesis. Activated by AGAMOUS in a cal-1, ap1-1 background. Expressed at low levels in developing stamens. Increased levels of ATH1 severely delay flowering in the C24 accession. Most remarkably, ectopically expressed ATH1 hardly had an effect on flowering time in the Col-0 and Ler accessions. ATH1 physically interacts with STM, BP and KNAT6 and enhances the shoot apical meristem defect of some of these genes suggesting a role in SAM maintenance. Nuclear localization is dependent upon interaction with STM. [URL-20]	Lee, J.-H. et al. (2008)

<b>Symbol</b>	<b>Full Name</b>	<b>Gene Model ID</b>	<b>Available Description</b>	<b>Literature</b>
<b>BEL1</b>	BELL 1	AT5G41410.1	[Arabidopsis thaliana] Homeodomain protein required for ovule identity. Loss of function mutations show homeotic conversion of integuments to carpels. Forms heterodimers with STM and KNAT1. Interacts with AG-SEP heterodimers is thought to restrict WUS expression. BEL interacts with MADS box dimers composed of SEP1(or SEP3) and AG, SHP1, SHP2 and STK. The interaction of BEL1 with AG-SEP3 is required for proper integument development and specification of integument identity. [URL-20]	Lee, J.-H. et al. (2008)
<b>BLH2 / SAW1</b>	BEL1-LIKE HOMEODOMAIN 2	AT4G36870.1	[Arabidopsis thaliana] Encodes a member of the BEL family of homeodomain proteins. Plants doubly mutant for saw1/saw2 (blh2/blh4) have serrated leaves. BP is expressed in the serrated leaves, therefore saw1/saw2 may act redundantly to repress BP in leaves. [URL-20]	Lee, J.-H. et al. (2008)
<b>BLH4 / SAW2</b>	BEL1-LIKE HOMEODOMAIN 4	AT2G23760.1	[Arabidopsis thaliana] Encodes a member of the BEL family of homeodomain proteins. Plants doubly mutant for saw1/saw2 (blh2/blh4) have serrated leaves. BP is expressed in the serrated leaves, therefore saw2 and saw1 may act redundantly to repress BP in leaves. [URL-20]	
<b>ChrGSP1</b>	GAMETE-SPECIFIC HOMEODOMAIN PROTEIN	AY052652.3	[Chlamydomonas reinhardtii] gamete-specific homeodomain protein	Lee, J.-H. et al. (2008)
<b>HvJUBEL</b>		AF334758.1	[Hordeum vulgare] homeodomain protein JUBEL1 (JuBel1) gene	
<b>PNY</b>	PENNYWISE	AT5G02030.1	[Arabidopsis thaliana] Mutant has additional lateral organs and phyllotaxy defect. Encodes a homeodomain transcription factor. Has sequence similarity to the Arabidopsis ovule development regulator Bell1. Binds directly to the AGAMOUS cis-regulatory element. Its localization to the nucleus is dependent on the coexpression of either STM or BP. [URL-20]	Lee, J.-H. et al. (2008)



<b>Symbol</b>	<b>Full Name</b>	<b>Gene Model ID</b>	<b>Available Description</b>	<b>Literature</b>
<b>YDA</b>	YODA	AT1G63700.1	[Arabidopsis thaliana] Member of MEKK subfamily, a component of the stomatal development regulatory pathway. Mutations in this locus result in embryo lethality. [URL-20]	Jeong, S.; Palmer, T.M.; Lukowitz, W. (2011)
<b>ZmKIP</b>	KNOTTED1-INTERACTING PROTEIN	AY082396.1	[Zea mays] knotted1-interacting protein (kip)	Lee, J.-H. et al. (2008)
<b><i>Cell cycle associated genes</i></b>				
<b>DP</b>	TRANSCRIPTION FACTOR DP	NM_001203285.1	transcription factor dp (DPB)	
<b>E2F</b>	E2F TRANSCRIPTION FACTOR 1	AT5G22220.2	[Arabidopsis thaliana] Member of the E2F transcription factors, (cell cycle genes), key components of the cyclin D/retinoblastoma/E2F pathway. Binds DPA and RBR1 proteins. Expressed throughout the cell cycle. Abundance increased by auxin through stabilization of the protein. Elevates CDK levels and activity, even under hormone-free conditions. Promotes cell division and shortens cell doubling time, inhibits cell growth. Transgenic plants overexpressing AtE2Fa contained an increased level of AtE2Fb transcripts that is paralleled by an increase in the amount of the AtE2Fb protein, suggesting that AtE2Fb expression might actually be up-regulated by the AtE2Fa transcription factor. [URL-20]	
<b>RBR</b>	RETINOBLASTOMA-RELATED	AT3G12280.1	[Arabidopsis thaliana] Encodes a retinoblastoma homologue RETINOBLASTOMA-RELATED protein (RBR or RBR1). RBR controls nuclear proliferation in the female gametophyte. Also required for correct differentiation of male gametophytic cell types. Regulates stem cell maintenance in Arabidopsis roots. Involved in the determination of cell cycle arrest in G1 phase after sucrose starvation. RBR1 is also involved in regulation of imprinted genes. Together with MSI1 it represses the expression of MET1. This in turn activates expression of the imprinted genes FIS2 and FWA. Functions as a positive regulator of the developmental switch	

Symbol	Full Name	Gene Model ID	Available Description	Literature
			from embryonic heterotrophic growth to autotrophic growth. [URL-20]	
<b>Auxin regulation</b>				
<b>ARF1</b>	AUXIN RESPONSE FACTOR 1	AT1G59750.1	[Arabidopsis thaliana] Encodes a member of the auxin response factor family. ARFs bind to the cis element 5'-TGTCTC-3' ARFs mediate changes in gene expression in response to auxin. ARF's form heterodimers with IAA/AUX genes. ARF1 enhances mutant phenotypes of ARF2 and may act with ARF2 to control aspects of maturation and senescence. ARF1:LUC and 3xHA:ARF1 proteins have a half-life of ~3-4 hours and their degradation is reduced by proteasome inhibitors. 3xHA:ARF1 degradation is not affected by a pre-treatment with IAA. A nuclear-targeted fusion protein containing the middle region of ARF1 linked to LUC:NLS has a similar half-life to the full-length ARF1:LUC construct. The degradation of 3xHA:ARF1 is not affected in an axr6-3 mutant grown at room temperature, although the degradation of AXR2/IAA7 is slowed under these conditions. [URL-20]	
<b>PIN1</b>	PIN-FORMED 1	AT1G73590.1	[Arabidopsis thaliana] Encodes an auxin efflux carrier involved in shoot and root development. It is involved in the maintenance of embryonic auxin gradients. Loss of function severely affects organ initiation, pin1 mutants are characterised by an inflorescence meristem that does not initiate any flowers, resulting in the formation of a naked inflorescence stem. PIN1 is involved in the determination of leaf shape by actively promoting development of leaf margin serrations. In roots, the protein mainly resides at the basal end of the vascular cells, but weak signals can be detected in the epidermis and the cortex. Expression levels and polarity of this auxin efflux carrier change during primordium development suggesting that cycles of auxin build-up and depletion accompany, and may direct, different stages of primordium development. PIN1 action on plant development does not strictly require function of PGP1 and PGP19 proteins. [URL-20]	

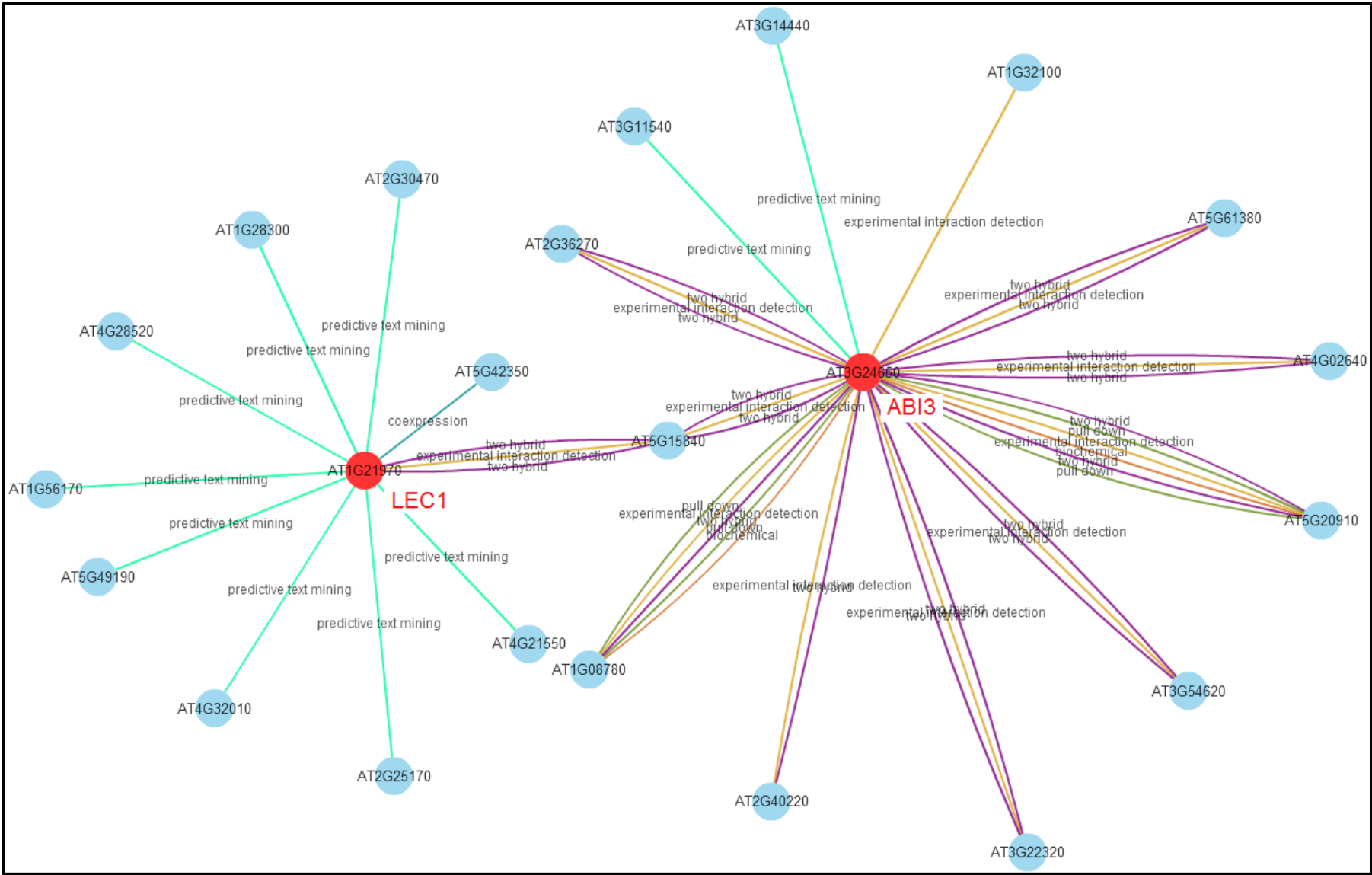
<u>Symbol</u>	<u>Full Name</u>	<u>Gene Model ID</u>	<u>Available Description</u>	<u>Literature</u>
<b>PIN2</b>	PIN-FORMED 2	AT5G57090.1	<p>[<i>Arabidopsis thaliana</i>]</p> <p>Encodes an auxin efflux carrier that is similar to bacterial membrane transporters. Root-specific role in the transport of auxin. Acts downstream of CTR1 and ethylene biosynthesis, in the same pathway as EIN2 and AUX1, and independent from EIN3 and EIN5/AIN1 pathway. In the root, the protein localizes apically in epidermal and lateral root cap cells and predominantly basally in cortical cells. Functions may be regulated by phosphorylation status. EIR1 expression is induced by brassinolide treatment in the brassinosteroid-insensitive br1 mutant. Gravistimulation resulted in asymmetric PIN2 distribution, with more protein degraded at the upper side of the gravistimulated root. Protein turnover is affected by the proteasome and by endosomal cycling. Plasma membrane-localized PIN proteins mediate a saturable efflux of auxin. PINs mediate auxin efflux from mammalian and yeast cells without needing additional plant-specific factors. The action of PINs in auxin efflux is distinct from PGP, rate-limiting, specific to auxins and sensitive to auxin transport inhibitors. Membrane sterol composition is essential for the acquisition of PIN2 polarity. [URL-20]</p>	
<b>PIN7</b>	PIN-FORMED 7	AT1G23080.1	<p>[<i>Arabidopsis thaliana</i>]</p> <p>Encodes a novel component of auxin efflux that is located apically in the basal cell and is involved during embryogenesis in setting up the apical-basal axis in the embryo. It is also involved in pattern specification during root development. In roots, it is expressed at lateral and basal membranes of provascular cells in the meristem and elongation zone, whereas in the columella cells it coincides with the PIN3 domain. Plasma membrane-localized PIN proteins mediate a saturable efflux of auxin. PINs mediate auxin efflux from mammalian and yeast cells without needing additional plant-specific factors. The action of PINs in auxin efflux is distinct from PGP, rate-limiting, specific to auxins and sensitive to auxin transport inhibitors. PINs are directly involved of in catalyzing cellular auxin efflux. [URL-20]</p>	

<u>Symbol</u>	<u>Full Name</u>	<u>Gene Model ID</u>	<u>Available Description</u>	<u>Literature</u>
<i>Gene silencing</i>				
<b>RH9</b>	RNA HELICASE 9	AT3G22310.1	[Arabidopsis thaliana] Sequence similarity to DEAD-box RNA helicases. Binds RNA and DNA. Involved in drought, salt and cold stress responses. [URL-20]	

## Appendix 2

<u>Symbol</u>	<u>Gene Model</u>	<u>Morex WGS-Contig</u>	<u>RNA-Seq-Contig</u>	<u>length</u>	<u>coverage 1+2</u>	<u>coverage 3+4</u>	<u>coverage 5+6</u>
ABI3	AT3G24650.1	morex_contig_56871	contig_137422	2622	13,758200	152,732647	172,943173
AGL38	AT1G65300.1	morex_contig_137013					
RH9	AT3G22310.1	morex_contig_37229	contig_108698	2314	38,321089	209,765774	116,061366
BBM	AT5G17430.1	morex_contig_44345	contig_31863	2895	0,275302	30,300518	448,015544
BLH2 / SAW1	AT4G36870.1	morex_contig_136249	contig_137432	2724	0,363069	1,407856	8,341410
BLH4 / SAW2	AT2G23760.1	morex_contig_136249	contig_137432	2724	0,363069	1,407856	8,341410
BP / KNAT1	AT4G08150.1	morex_contig_1561605	contig_34186	1476	0,000000	0,210705	7,731030
EMK / AIL5	AT5G57390.1	morex_contig_44345	contig_31863	2895	0,275302	30,300518	448,015544
LEC1	AT1G21970.1	morex_contig_2547173	contig_31374	1390	0,329496	3,542446	344,974101
MYB115	AT5G40360.1	morex_contig_42106	contig_109134	1538	7,885566	35,983095	102,119636
MYB118	AT3G27785.1	morex_contig_42106	contig_109134	1538	7,885566	35,983095	102,119636
OsH15	OS_P46609.2	morex_contig_1561605	contig_34186	1476	0,000000	0,210705	7,731030
PIN2	AT5G57090.1	morex_contig_50370	contig_104300	390	0,000000	0,969231	5,738462
			contig_139725	661	302,945537	228,243570	219,582451
			contig_147899	300	0,000000	0,000000	2,786667
			contig_49920	378	0,000000	0,000000	4,682540
TaRKD1	TA_AEB26835.1	morex_contig_65307	contig_76112	1613	1,905146	164,810911	490,606944
RKD4 / GRD	AT5G53040.1	morex_contig_495	contig_76112	1613	1,905146	164,810911	490,606944
STM	AT1G62360.1	morex_contig_1561605	contig_34186	1476	0,000000	0,210705	7,731030
WOX2	AT5G59340.1	morex_contig_47071	contig_85409	1222	0,147300	4,940262	62,695581
WOX8	AT5G45980.1	morex_contig_6024	contig_32577	1972	0,000000	0,454361	41,595335
WOX9	AT2G33880.1	morex_contig_140569	contig_33893	2047	0,000000	1,388373	58,441133
WUS	AT2G17950.1	morex_contig_1585113	contig_85409	1222	0,147300	4,940262	62,695581

Appendix 3



## Appendix 4

\*\*\*

### README

**In this file all used tools are described. It can be read where the tool can be attained, which settings have to be adjusted or which further important instructions for use have to be followed. If required an installation guideline is also written down.**

### BLASTN

BLAST is the abbreviation for Basic Local Alignment Search Tool and is one of the most used tools in computational biology. First intention is to find homologous sequences and identify species. Second aim is to locate functional domains, which can be achieved by working with protein sequences. The third intention is to establish phylogeny between species on sequence scaffolds. Second to last DNA mapping is an important function, which provide the comparison of query sequences with physical chromosomal positions and it can be used to find unknown locations of genes and functional sites. At last the algorithm can be used to map annotations from a well-known organism to an unknown.

### WEBSITE

<http://webblast.ipk-gatersleben.de/barley/>

### INSTALL // INPUT

Query: Nucleotide sequence (cDNA) of a RNA-Seq-Contig in Fasta format.

Database: assembly\_WGSMorex.

[Whole genome sequence assembly of the Morex cultivar (Set of genomic nucleotide sequences).].

Settings: Default.

### NOTICE

The IPK Barley BLAST Server provides also further databases (several cultivars) and can be used for other types of BLAST applications (e.g. BLASTx, BLASTp,...).

**BLASTX**

The BLASTx algorithm translates a nucleotide query sequence into a protein sequence and searches, with this translated sequence, against given protein subject sequences or a protein database.

**WEBSITE**

[http://blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastx&BLAST\\_PROGRAMS=blastx&PAGE\\_TYPE=BlastSearch&SHOW\\_DEFAULTS=on&LINK\\_LOC=blasthome](http://blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastx&BLAST_PROGRAMS=blastx&PAGE_TYPE=BlastSearch&SHOW_DEFAULTS=on&LINK_LOC=blasthome)

**INSTALL // INPUT**

Query: Nucleotide sequence (cDNA) of a RNA-Seq-Contig in Fasta format.

Database: Non-redundant protein sequences (nr).

Settings: Default.

**NOTICE**

For the analysis only the query sequence needed to be inserted into the input window and the 'BLAST' button has to be clicked. This starts the search run immediately.

**CLUSTALW2**

ClustalW2 produces biological significant multiple sequence alignments (MSA) for a set of DNA sequences or protein sequences. It calculates the most likely MSA in three basic steps. As the first step all sequences of the input set were aligned pairwise (global) and arranged due to their alignment score. Consequential a distance matrix is formed with the calculated arrangement in the second step. Finally a MSA is produced out of the distance matrix. The main intention in generating a MSA is to identify conserved sequence regions, sites or domains in the input sequence set. This can be achieved by including known annotated sequences into the analyses. The second general aim is to show evolutionary relationships and shared lineages between various species. The phylogenetic relation can be shown with the Cladograms or Phylograms ClustalW2 produces automatically beside the MSA.

**WEBSITE**

<http://www.ebi.ac.uk/Tools/msa/clustalw2/>



---

**INSTALL // INPUT**

STEP 1 - Query: File with similar protein sequences from related organisms in Fasta format

STEP 2 - Pairwise Alignment Options:   GAP OPEN → 100  
  GAP EXTENSION → 10.0

STEP 3 - MSA Options:   GAP OPEN → 100  
                          GAP EXTENSION → 10.0  
                          ORDER → input

STEP 4 - Submit: Starting the analysis.

**NOTICE**

After the job has been done the results are displayed. At first the alignment is presented and with clicking onto the 'Download Alignment File'-button the alignment can be stored locally. This file can be observed with e.g. the Editor or WordPad. Using the next tab named 'Result Summary' there is on the right site a rectangle with the headline 'JalView'. Beneath this there is a button which starts the JalViewer which provides the MSA visualization. Sometimes there is an additional window which opens automatically with clicking on the 'Start JalView'-button. There you have to allow the program to access to the computer. Subsequently the tool opens itself and you can scroll to see the whole alignment and in the menu bar you can test the color options and further applications.

**CDD**

The Conserved Domain Database (CDD) is a protein annotation resource that consists of well-annotated multiple sequence alignment models for protein domains and full-length proteins. The CDD mainly consists of domains curated by the NCBI. These domains are annotated with 3D structures, defined domain boundaries and an insight into sequence-structure-function relationship if present. Further content of the database are domain models from external source databases like Pfam, SMART, COG or TIGRFAM. The main properties of the database are to visualize the architecture of protein domains, highlight the presence of domains on the sequence, identify a putative function of an unknown protein sequence and eventually identify specific amino acids in a sequence that are putatively involved in functions as DNA binding or catalysis.

**WEBSITE**

[http://www.ncbi.nlm.nih.gov/Structure/cdd/docs/cdd\\_search.html](http://www.ncbi.nlm.nih.gov/Structure/cdd/docs/cdd_search.html)

**INSTALL // INPUT**

Query: Nucleotide sequence (cDNA) of a RNA-Seq-Contig in Fasta format.

---

Settings: Default.

## MULAN

Mulan is the abbreviation for multiple-sequence local alignment and is a new integrative comparative application that generates textual and particularly graphical multiple-sequence local alignments (MSLAs). The tool Mulan produces rapid, dynamical and very accurate local alignments for both closely and distantly related organisms. Especially for distant organisms Mulan ensures a reliable representation of short- and large-scale genomic rearrangements. For the graphical visualization of the finished MSLAs there are different options for achieving a suitable presentation, e.g. the reference sequence of the pairwise alignments can be flexibly be changed.

## WEBSITE

<http://mulan.dcode.org/>

## INSTALL // INPUT

STEP 1 - Enter the number of sequences, you want to align against each other, in the box of 'NUMBER OF SPECIES'.

STEP 2 - Click on the 'Select'-button on the left site beneath 'ALL FINISHED SEQUENCES'.

STEP 3 - Paste or upload your sequences into the text boxes for the sequences. And submit it.

STEP 4 - If all sequences could be aligned pairwise you have a first look onto a phylogenetic tree. On the right site there is a red button with '>> Continue' which submit the tree to the final alignment step. Click it.

STEP 5 - Click onto the link 'Dynamic visualization' and the multiple local sequence alignment is displayed in another window.

STEP 6 - Choose the settings:

- ECR similarity at least → 50 %
- Graph height → 200 pixels
- REFRESH

## NOTICE

The picture settings needed to be optimized as required for the example. If you click onto the Request ID in top left the previous GUI is displayed again. Beneath there is the MultiTF button which can be used to identify all available TFBS on the query sequences.

## MULTIF

STEP 1 - Checkmark the 'plant' TRANSFAC library and submit the query.

STEP 2 - Click on the 'SELECT ALL'-button and submit the query again.

STEP 3 - If the data was successfully submitted you need to click the 'CHECK IT'-button.

STEP 4 - Click on the small picture at the right for opening the visualization.

STEP 5 - Choose the settings:     Picture: Bases per layer → 20kb

   Picture width: 800

   Show binding sites: multi-species AND all

SUBMIT

## **TRIANNOT**

TriAnnot is a pipeline for the automated structural and functional annotation of plant genomes, which can be accessed through a web interface for small scale analysis. The pipeline has a modular architecture allowing a simultaneous annotation of protein-coding genes, identification of conserved non-coding sequences and detection of molecular markers. TriAnnot combines methods and applications from other pipelines with the intention to integrate the most innovative features of these already available pipelines.

### **WEBSITE**

<http://wheat-urgi.versailles.inra.fr/Tools/Triannot-Pipeline>

### **INSTALL // INPUT**

STEP 1 - Ask an account.

STEP 2 - Run Pipeline. Enter user name and password. Click 'OK'.

STEP 3 - Enter an analysis title. Choose the template → 'Barley default analysis'

STEP 4 - Enter a query sequence in Fasta format between 10 Kb and 3 Mb. Submit analysis.

STEP 5 - If you have an eMail with the notice of the finished analysis click on the tab 'My Analysis'.

STEP 6 - In the column 'Status' click on the rightmost icon to view the results with GBrowse.

STEP 7 - Click on the tab 'Select Tracks' and choose all features you want to visualize in the browser.

STEP 8 - Click on the tab 'Browser' to observe the selected features.

### **NOTICE**

Unfortunately the input sequences have to be between 10 Kb and 3 Mb sequence length. If this is not the case you can cheat by adding Ns on your query sequence. Just add a sequence of Ns with the missing number of bases onto your sequence. You can distinguish the 'cheat sequence' to the 'correct sequence' by the GC content. Because the GC content directly collapse if there are only Ns in the sequence.

**TABLET**

Tablet is a software for the visualization of next-generation sequence assemblies and alignments, which is freely available and employable for users of all abilities. Tablet supports most of the common input assembly formats and is able to process a large number of reads. The interface of Tablet provides an overview of the whole selected contig and therefore gives access to any region of the assembly with an intuitive navigation. All regions of the assemblies are displayed with high-quality at any zoom level.

**WEBSITE**

<http://bioinf.scri.ac.uk/tablet/>

**INSTALL // INPUT**

STEP 1 - Use the following link: <http://bioinf.scri.ac.uk/tablet/download.shtml>.

STEP 2 - Download Tablet most suitable for your operating system.

STEP 3 - Install the downloaded package on your PC with the default instructions.

STEP 4 - Open the program.

STEP 5 - Click the "Open Assembly"-Button. Two input fields will open.

STEP 6 - "Primary assembly file or URL:" → Put a mapping file like SAM or BAM here.

STEP 7 - "Reference/consensus file or URL": → Here you need to enter the reference sequence which was also used in the prior mentioned mapping in FASTA or FASTQ format.

**NOTICE**

The main effort in using tablet is generating the primary assembly and the corresponding reference sequence. But these assembly and reference files can be provided by some members of the BIT group at the IPK. The larger the mapping files the more memory usage is required and therefore more patience is to be exercised.

\*\*\*